Combined approach for terminology extraction: lexical statistics and linguistic filtering

Béatrice Daille
TALANA
University Paris 7
Case 7003
2, Place Jussieu
F-75251 Paris Cedex 05
France
daille@linguist.jussieu.fr

March 1994

Abstract

This paper describes the automatic extraction of the terminology of a specific domain from a large corpus. The use of statistical methods yields a number of solutions, but these produce a considerable amount of noise. The task we have concentrated on is the creation and testing of an original method to reduce high noise rates by combining linguistic data and statistical methods. Starting from a rigorous linguistic study of terms in the domain of telecommunications, we designed a number of filters which enable one to obtain a first selection of sequences that may be considered as terms on the grounds of morphosyntactic criteria. Various statistical methods are applied to this selection and the results are evaluated. The best statistical model - that is to say, the one that gives a correct list of terms with the lowest rates of noise and silence - turns out to be the one based on the likelihood ratio - in which frequency is taken into account. This result contradicts numerous previous results regarding the extraction of lexical resources, which claim that the association ratio (for example the use of mutual information) is more significant than frequency.

Contents

1 Methodological Principles **3**

2 Linguistic Extraction and Frequency Count **4**
- 2.1 General Concepts 4
  - 2.1.1 Regular Languages 4
  - 2.1.2 Finite automaton languages 4
- 2.2 Definition: terms as co-occurrences 5
- 2.3 Finite automata for terms of length 2 10
  - 2.3.1 N_1 N_2 10
  - 2.3.2 N_1 PREP (DET) N_2 11
  - 2.3.3 N ADJ 13
- 2.4 Program description 14
- 2.5 Program results 15

3 Statistical scores **18**
- 3.1 Frequencies 19
- 3.2 Association criteria 19
- 3.3 Shannon diversity 22
- 3.4 Distance Measures 23
- 3.5 Affinity score 24

4 Graphical evaluation of statistics **24**
- 4.1 Reference list 24
- 4.2 Graphical evaluation 25

5 Examination of the results **33**
- 5.1 Frequency 34
- 5.2 Association criteria 36
  - 5.2.1 Associatio ratio and Cubic association ratio 36
  - 5.2.2 Log-Likehood coefficient 37
  - 5.2.3 Fager and MacGowan coefficient 38
- 5.3 Diversity 40
- 5.4 Distance Measures 42

6 Conclusion **45**

A French Tags **50**

B Sorting of the pairs obtained with the log-likehihood coefficient **53**

We present here the methodology that we have chosen to extract terms from a corpus by combining statistical scores and linguistic data. In the first part of this paper, we develop our methodology and define our goal. We take into account linguistic specifications of terms established in [Daille, 1994] and decide which terms seem to fit a statistical approach.
In a second part, we describe the program which extracts and counts term co-occurrences: this program uses patterns which are encoded in finite automata. After a short description of this program, we examine the extracted co-occurrences and comment upon the results obtained.
In the third part, we present the statistical scores that we have tested and the graphic evaluation that suggests that only a few of them should be retained. In the fourth part, we examine the way terms are sorted by each of the retained scores and other informations provided by Shannon diversity and distance measures. Only one criterion will be retained.

1 Methodological Principles

The objective of our work is to define a methodology for automating the extraction of terms using corpora and statistical scores. This method is defined for monolingual corpora, but it can be extended to bilingual terminology as shown in [Daille *et al.*, 1994].
Statistical methods play an important part in NLP and give good results in several fields such as part-of-speech tagging or sentence alignment. In the field of the extraction of monolingual lexical resources, the works of [Lafon, 1984] on French, [Calzolari and Bindi, 1990] on Italian and numerous ones on English such as [Church and Hanks, 1990] have been reported. Statistical scores applied to corpora produce quantitative information on the lexical affinities that words share. Some of these lexical affinities are co-occurrences. The problem with the co-occurrences obtained by any purely statistical method is their tremendous diversity. If we refer to [Lafon, 1984], the extracted co-occurrences represent either semantic associations, or functional associations where terms are found, and miscellaneous associations. Our goal is to extract terms and to disregard other types of co-occurrences. The problem is twofold:

- Firstly, to guide the statistical score on co-occurrences that we wish to extract. We have decided to extract terms of length 2; the length of a multi-word-unit (MWU) is defined as the number of *main items* it contains[1]. Indeed, terms of length 2 are by far the most frequent ones. Moreover, the majority of terms whose length is greater than 2 are combinaisons of terms of length 1 or 2. To use a statistical score, sampling is necessary; terms of length 2 seem sufficiently represented in our corpus to allow a statistical approach. To detect them, we use their linguistic specifications in terms of morphosyntactic structures (patterns). These patterns are described by regular expressions and can thus be

[1] *Main items* are nouns, adjectives, adverbs, etc. Neither prepositions nor determiners are main items

extracted from the tagged corpus with finite automata. The extracted co-occurrences are all terms of length 2 permissible, from a morphosyntactic point of view.

- Secondly, to use statistical scores in order to determine which of these co-occurrences are true terms. We will examine various statistical scores; scores that have been already used in NLP, but also scores that have been used in other scientific areas such as biology. We will evaluate these scores to discover which scores are the best for our purpose: our aim is to select a score that assigns high values to terms of our technical field and low values to co-occurrences which are not technical terms.

We now present how we have decided to guide the statistical scores.

2 Linguistic Extraction and Frequency Count

The finite state automata that we have produced detect only terms of length 2: this point is justified linguistically and statistically. We will briefly present the general concept of finite state automata with two definitions: regular expressions and finite automata. We will then describe the program and its output.

2.1 General Concepts

2.1.1 Regular Languages

Given an alphabet A:

1. $\emptyset$ is a regular language,
2. For any string $x \in A^*$, x is a regular language,
3. If X and Y are regular languages, so is $X \cup Y$,
4. If X and Y are regular languages, so is the concatenation of X and Y, denoted XY,
5. If X is a regular language, so is X^*. Given a set of strings X, the Kleene star or closure of X, denoted X^*, is the set formed by concatening members of X together any number of times (including zero) in any order and allowing repetitions,
6. Nothing else is a regular language unless its being so follows from 1-5.

2.1.2 Finite automaton languages

Finite automaton languages, like regular languages, allow one to describe Kleene languages. A finite automaton language is defined by the quintuple: (A, E, IF, F, T) where A denotes its alphabet, E its set of states, IF its set of initial states, F its set of final states, and T its set of transition functions. A transition

function is defined by the 3-tuple: (e_i, e_j, l) where e_i and e_j are states and l belongs to the alphabet. A finite state automaton is an abstract computing device that receives a string of symbols as input, reads the string one symbol at a time from left to right and after reading the last symbol halts and signifies either acceptance or rejection of the input. At any point of its computation a finite automaton is in one of a finite number of states. The computations of a finite automaton are directed by a program, which is a finite set of instructions for changing from state to state as the automaton reads input symbols. A computation always begins in a designated state, the initial state. There is also a specified set of final states; if the finite automaton ends up in one of these after reading the input, it is accepted; otherwise, it is rejected. Having characterized regular languages and finite automaton languages, we use the Kleene theorem to show that they are in fact identical:

> A set of strings is a finite automaton language if and only if it is a regular language.

Finite automaton languages and regular languages are equivalent. We will describe our patterns with finite automaton diagrams for reasons of readability.

2.2 Definition: terms as co-occurrences

A co-occurrence is an association of two lexical units found in a corpus. Terms of length 2 are co-occurrences with the following properties: (a) they are described thanks to their morphosyntactic structure, (b) they accept modifications that could create new terms whose length will be greater than 2, and (c) they accept variants. A co-occurrence which characterizes a term of length 2 has the following properties:

1. It is oriented and follows the linear order of the text,
2. It is composed of two main items,
3. It must belong to the morphosyntactic structure set that characterize terms of length 2.

Determining these properties implies a preliminary treatment of the corpus: all the lexical units must receive a part-of-speech tag. Each extracted co-occurrence characterizes a pair: this pair is composed of the two lemmas which are the two main items of a specific pattern. In order to obtain pairs that are as general as possible, we use lemmas instead of full-form items. This choice requires the association of each item in the corpus with its lemma. Each co-occurrence is counted equally: we consider that each term appears equiprobably in the corpus. We have decided to extract terms of length 2 that we will call *base-term* which matched a list of previously determined patterns:

N ADJ *station terrienne* (*Earth station*)

$\mathbf{N}_1$ *de* (DET) $\mathbf{N}_2$ *zone de couverture* (*coverage zone*)

$\mathbf{N}_1$ *à* (DET) $\mathbf{N}_2$ *réflecteur à grille* (*grid reflector*)

N_1 PREP N_2 *liaison par satellite* (*satellite link*)

N_1 N_2 *diode tunnel* (*tunnel diode*)

Of course, terms exist whose length is greater than 2. But the majority of terms of length greater than 2 are created recursively from base-terms. We have distinguished three operations that lead to a term of length 3 from a term of length 1 or 2: "overcomposition", modification and coordination. We illustrate now these operations with a few examples where the base-terms appear inside brackets:

1. **Overcomposition**
   Two kinds of overcomposition have been pointed out: overcomposition by juxtaposition and overcomposition by substitution.

   (a) Juxtaposition
   A term obtained by juxtaposition is built with at least one base-term whose structure will not be altered. The example below illustrate the juxtaposition of a base-term and a simple noun:
   N_1 $PREP_1$ [N_2 $PREP_2$ N_3]
   modulation par [déplacement de phase] ([phase shift] keying)

   (b) Substitution
   Giving a base-term, one of its main item is substituted by a base-term whose head is this main item. For example, in the N_1 $PREP_1$ N_2 structure, N_1 is subtituted by a base-term of N_1 $PREP_2$ N_3 structure to create a term of N_1 $PREP_2$ N_3 $PREP_1$ N_2 structure:
   réseau à satellites + *réseau de transit* → *réseau de transit à satellites* (*satellite transit network*).
   We notice in the above example that the structure of *réseau à satellites* (*satellite network*) is altered.

2. **Modification**
   Modifiers that could generate a new term from a base-term appear either inside or after it.

   (a) Insertion of modifiers
   Adjectives and adverbs are the current modifiers that could be inserted inside a base-term structure: adjectives in the N_1 PREP (DET) N_2 structure and adverbs in the N ADJ one:
   liaisons **multiples** *par satellite* (*multiple [satellite links]*)
   réseaux **entièrement** *numériques* (*all [digital networks]*)

   (b) Post-modification
   Adjectives and adverbial prepositionnal phrase of PREP ADJ N are the main modifiers that lead to the creation of new terms: post-adjectives can modify any kind of base-terms; for example, *[station terrienne] brouilleuse* (*interfering [earth(-)station]*).
   Adverbial prepositional phrases modify either simple nouns or base-terms[2]: *amplificateur(s) [à faible bruit]* (*[low noise] amplifier(s)*),

[2]In this case, the length of the term is equal to 4

[interface(s) usager-réseau] [à usage multiple] (*[multipurpose] [user-network interface(s)]*).

3. **Coordination**
   Coordination is a rather complex syntactic phenomenon (term coordination have been studied in [Jacquemin, 1991]) and seldom generates new terms. Let us examine a rare example of a term of length 3 obtained by coordination :
   N_1 *de* N_3 + N_2 *de* N_3 $\rightarrow$ N_1 *et* N_2 *de* N_3
   assemblage de paquet + *désassemblage de paquets* $\rightarrow$ *assemblage et désassemblage de paquets* (*packet assembly/desassembly*)

Now do we have to take into account modifications that affect the structure of terms in our frequency counts? These counts are crucial as they are the parameters of the statistical scores. An incorrect frequency count produces incorrect or irrelevant values by statistical scores. This choice is difficult and important. This decision relies upon the nature of the modification that affects the term structure. Let us examine how we handle the counts of the co-occurrences following each particular case:

1. Terms of length 2:
   If the sequence *antenne de réception* (*receiving antenna*) is encountered, the number of occurrences of the pair (antenne, réception) is incremented by 1; if the sequence *réception de l'antenne* (*antenna reception*) is encountered, the number of occurrences of the pair (réception, antenne) is incremented by 1.

2. Terms of length 2 composed, modified or coordinated. Let us examine each case:

   (a) Composition.
   Composition is not taken into account. Thus, if we encounter the sequence: *antenne de réception du satellite* (*satellite receiving antenna*) associated with the pattern: N_1 PREP N_2 PREP N_3, we count the occurrences of the pairs (antenne, réception) and (réception, satellite), but do not count the occurrence of the pair (antenne, satellite). This decision is motivated by, on one hand, the fact that we want first to extract terms of length 2, and on the other hand, if we plot the occurrences of the pairs (antenne, réception) and (antenne, satellite), we alter the frequency count: for example, only one occurrence of *antenne* will lead to two occurrences, namely one for the pair (antenne, réception) and the other for the pair (antenne, satellite). In order not to alter the frequency count and to avoid artificially increasing the number of occurrences of some pairs, we do not take composition into account.

   (b) Post-modification
   Post-modification is also not taken into account. Let us examine the

sequence *bande passante étroite* (*narrow passband*) which is either a composed term built with two terms of length 2, *bande passante* (*passband*) and *bande étroite* (*narrow band*), or the term of length 2 *bande passante* (*passband*) post-modified by the adjective *étroite* (*narrow*); *passante* is not considered as a adjective modifier of *bande étroite* as it is not possible to insert an adjective inside a term of structure N ADJ in French. As we do not know the status of the sequences *bande passante* and *bande étroite*, we consider that the adjective *étroite* post-modifies the sequence *bande passante*. Thus, we plot only the occurrence of the term: the number of occurrences of the pair (bande, passante) is incremented by 1. In the same way, if we meet the sequence *antenne de réception parabolique* (*parabolic receiving antenna*), we plot only the occurrence of the pair (antenne, réception).

(c) Insertion of modifier.
The insertion of an adjective modifier inside the structure N_1 PREP (DET) N_2 poses problems because it is very frequent in French. It is not possible to ignore it, but if we take it into account, the frequency count will be altered: with the sequence *antenne parabolique de réception* (*parabolic receiving antenna*), the number of occurrences of the pairs (antenne, parabolique) and (antenne, réception) is incremented by 1. We have chosen to proceed in two steps to extract terms of length 2: first, we extract terms of structure N_1 (PREP (DET)) N_2 separately from terms of structure N ADJ. So, for the sequence above, the number of occurrences of the pair (antenne, parabolique) is incremented by 1 when we count the occurrences of the term of structure N ADJ, and the occurrences of the pair (antenne, réception) is incremented by 1 when we count term of structure N_1 (PREP (DET)) N_2. Having accepted that the structure N_1 PREP (DET) N_2 could be modified by an inserted adjective implies that terms of length 3 of structure N_1 ADJ PREP (DET) N_2 are considered as two terms of length 2 of structures N ADJ and N_1 PREP (DET) N_2. However, we will see that statistical scores allow one to differentiate terms of length 3 from terms of length 2 modified by an inserted adjective. For example, the pair (service, satellite) appears in our list of term of length 2 of structure N_1 (PREP (DET)) N_2, but the program indicates that the pair (service, satellite) appears most often under the form *service fixe par satellite* (*fixed-satellite service*). We can then deduce that *service fixe par satellite* is a term of length 3 and that *service fixe* is not a term of length 2. Furthermore, we must point out that the case described for (service, satellite) and (service, fixe) is rarely encountered. The most frequent case is the one illustrated by the sequence *antenne parabolique de réception* where the pair (antenne, réception) appears most often in the form *antenne de réception* and not in the form *antenne parabolique de réception*. A decision must be made on the status of the sequence *antenne parabolique de réception*

(*parabolic receiving antenna*), namely composed-term with the two terms of length 2 *antenne parabolique* (*parabolic antenna*) and *antenne de réception* (*receiving antenna*), or modified-term where the term *antenne de réception* is modified by the adjective *parabolique* is difficult. With regards to this problem, we consider *antenne de réception* and *antenne parabolique* as two candidate terms without making a choice.

We are aware that the decisions that we have taken are not the only possible ones. Another treatment could have been the following: to record the N ADJ morphosyntactic sequence under the corresponding pair only if this sequence is not a sub-sequence of the N_1 (PREP (DET)) N_2 pattern. But proceeding in such way would require the recognition of many composed-terms and the miss of many terms of length 2. For example, for the sequence *antenne parabolique du satellite* (*parabolic antenna of the satellite*), the occurrence of the pair (antenne, satellite) would be plotted for the N_1 (PREP (DET)) N_2 pattern and the occurrence of the pair (antenne, parabolique) would not be plotted for the N ADJ pattern. It is because we desire to extract term of length 2 that we distinguish between N_1 (PREP (DET)) N_2 and N ADJ patterns. This method has certain drawbacks of which the most important is the non-presence of the N ADJ terms in an unique conceptual sort.

In general, other types of inserted modifiers are accepted if they do not falsify the frequency count, as the adverb in the N ADJ structure. The allowed inserted modifiers are specified in section 2.3.

(d) Coordination

Let us consider the sequence *équipements de modulation et de démodulation* (*modulation and demodulation equipments*). The number of occurrences of the pairs (équipement, modulation) and (équipement, démodulation) are both incremented by 1. In this case, even if *équipement* appears only once in the text, it seems reasonable to consider that it appears twice because we could have encountered the sequence *équipement de modulation et équipement de démodulation* (*modulation equipment and demodulation equipment*). This frequency calculation is incorrect only if the N_1 PREP N_2 CONJ (PREP) N_3 structure, where CONJ is a coordination conjunction, is a term of length 3. However, since these terms are extremely rare, we accept this margin of error.

Now, we present techniques that allow us to count only co-occurrences that we wish to take into account. We have chosen to use finite automata. The occurrences of the extracted sequences are classified as a pair: a pair is oriented, composed with two lemmas and collected all the sequences where the full form of the two lemmas appear either in the N_1 (PREP (DET)) N_2 or N ADJ pattern. It is to these pairs that the statistical scores we will introduce in section 3 apply.

2.3 Finite automata for terms of length 2

It is for good reason that we decided to use finite automata: indeed, most of the programs that extract co-occurrences use some form of window which scans the corpus. There are major problems with the window technique: the window's size is arbitrary and it is not possible to filter the morphosyntactic structures. The use of rigorous linguistic filters is necessary for the frequency plot.
The finite automata written to count and extract morphosyntactic sequences that characterize terms of length 2 are not determinist. Several initial states and final states are possible. These finite automata have been updated with our tagged and lemmatized corpus and can only be used on corpora using this tagset (the French tagset is given in Annex A). These finite state automata take into account the behavior of the IBM tagger; if these finite automata are used with the same tagset but with a different tagger, they should be modified on several points as will be described. The alphabet of our finite automata is a subset of the tagset. We added symbols to denote lemmas, full-form items, and graphic codes. We also integrated in these finite automata a procedure to check the agreement in gender and number.
The finite automata are considered as linguistic filters and are used to count correctly the frequency of co-occurrences. Finite automata are represented by state diagrams where:

- the initial state is indicated by an arrow,
- the final state is indicated by a double square,
- states are squares and transition functions are indicated by lines leading from the right side of a state to the left side of another state. For the sake of readability, the tag that must be read to reach a state is indicated on the reached state,
- the recorded lemma pairs correspond to the states shown with a bold frame; the states shown with a dotted bold frame correspond to intermediary final states; the occurrence of a pair of which one of the lemma is associated to an intermediary final state is recorded only if a final state is reached afterwards.

These finite automata have been modified several times in order to optimize the pair extraction: indeed, we managed to minimize the number of bad candidates by prohibiting several patterns that lead more bad candidates than good ones. Finite automata for English were also written for terms of length 2 (N_2 N_1, ADJ N, N_1 PREP N_2); these will not be described further. However, results produced by English finite automata will be used in our program of bilingual terminology extraction (see [Daille *et al.*, 1994]).

2.3.1 N_1 N_2 finite automaton

Figure 1 shows the finite automaton used to identify terms of structure N_1 N_2; it is indicated in the figure that neither of the two lemmas are written in

upper case: this condition has been stipulated to eliminate abbreviations. The presence of an abbreviation implies a composed-term and our goal is to isolate only terms of length 2. In the other figures showing other types of term, this condition is no longer illustrated but is still required. Unfortunately, this lexical restriction which is imposed upon nouns does not allow the elimination of all the abbreviations as some of them are in lower case. This finite automaton is

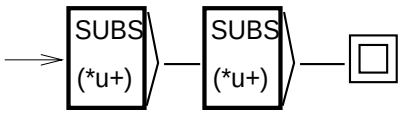

Figure 1: N_1 N_2

very simple: terms of structure N_1 N_2 are the most frozen ones. In our corpus, it appears that their structures are never modified. This frozen syntactic structure is not necessarily found in other technical domains.

2.3.2 N_1 PREP(DET) N_2 finite automata

For sake of readability, we present separately the finite automata for terms of length 2 of structure N_1 *de* (DET) N_2, N_1 *à* (DET) N_2 and N_1 PREP N_2. In figures 2 and 3, coordinations that have been taken into account are not represented. These coordinations are:

- right coordinations with comma such as: *circuits de commande, d'affichage et d'alarme* (*control, alarm and display circuits*); these coordinations must share the following properties:
  - the nouns coordinated with a comma accompany the preposition *de* (resp. *à*), the determinant being optional: a sequence such as *circuits de commande, affichage et alarme* is not recognized but the finite automaton,
  - the last noun of the sequence should coordinated to others thanks to a coordinate conjunction (i.e. not a comma): a sequence as *circuits de commande, d'affichage, d'alarme* is not recognized,
- left coordination with coordinate conjunction (i.e. not a comma) as in the following sequence: *convertisseurs-élévateurs et abaisseurs de fréquence* (*up and down converters*); only sequences which are preceded by an article or a contracted preposition are recognized.

These restrictions have been introduced in order not to take into account sequences which are not coordinations. Coordinations which use terms of different structures are not treated. A optional determinant is allowed for these two patterns between the preposition *de* (resp. *à*) and N_2.

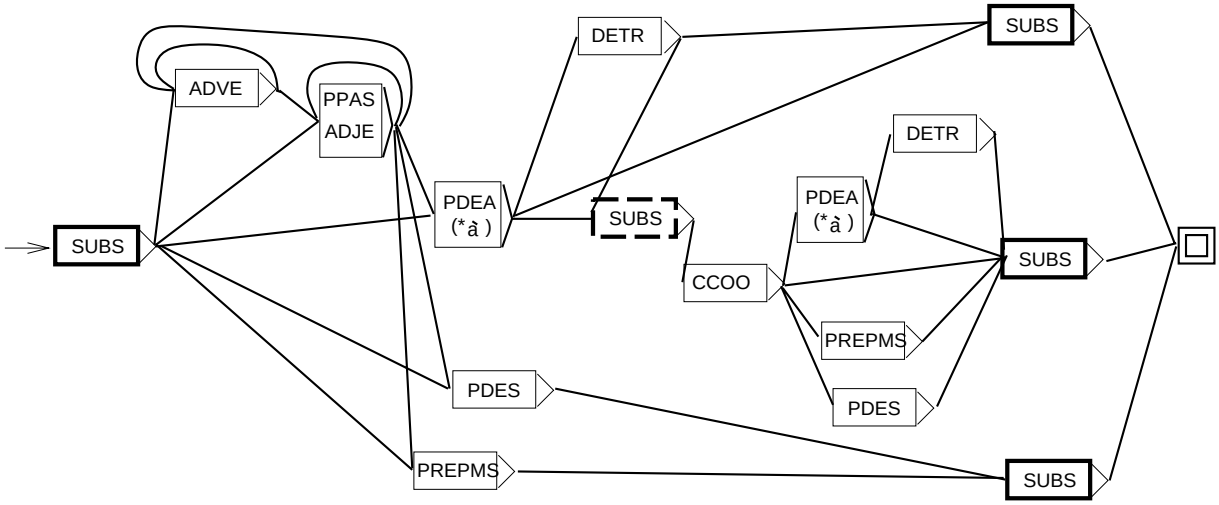

Figure 2: N_1 de (DET) N_2

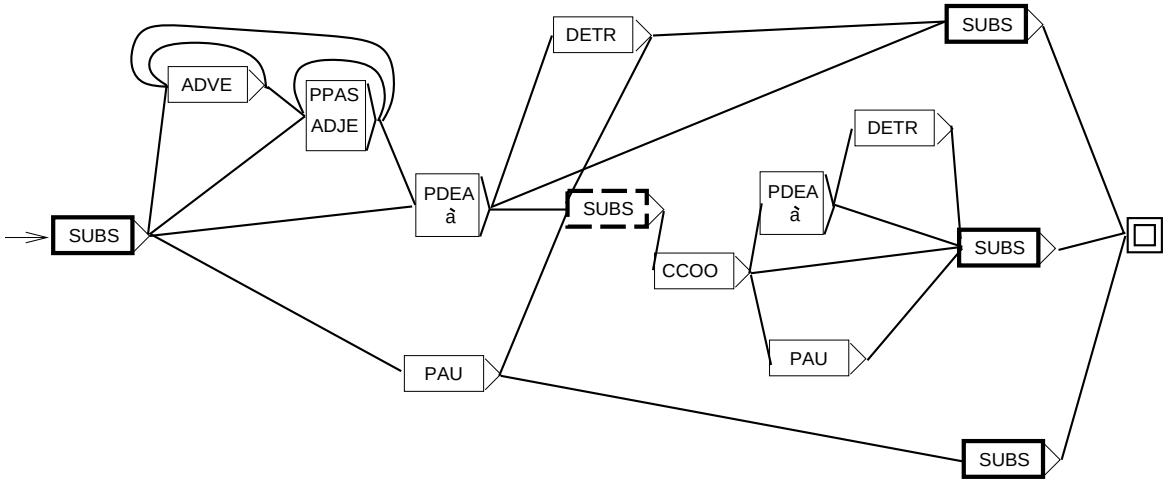

Figure 3: N_1 à (DET) N_2

For the N_1 PREP N_2 pattern where PREP is neither the preposition *à*, nor the preposition *de*, no optional determinant is allowed. We have observed that we obtain many more bad candidates if an optional determinant was allowed for this pattern.

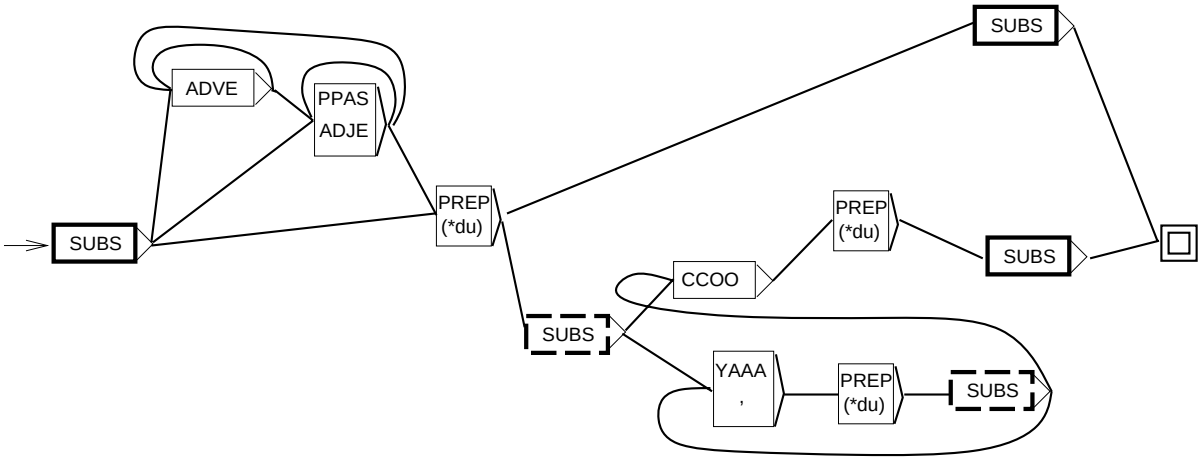

Figure 4: N_1 PREP N_2

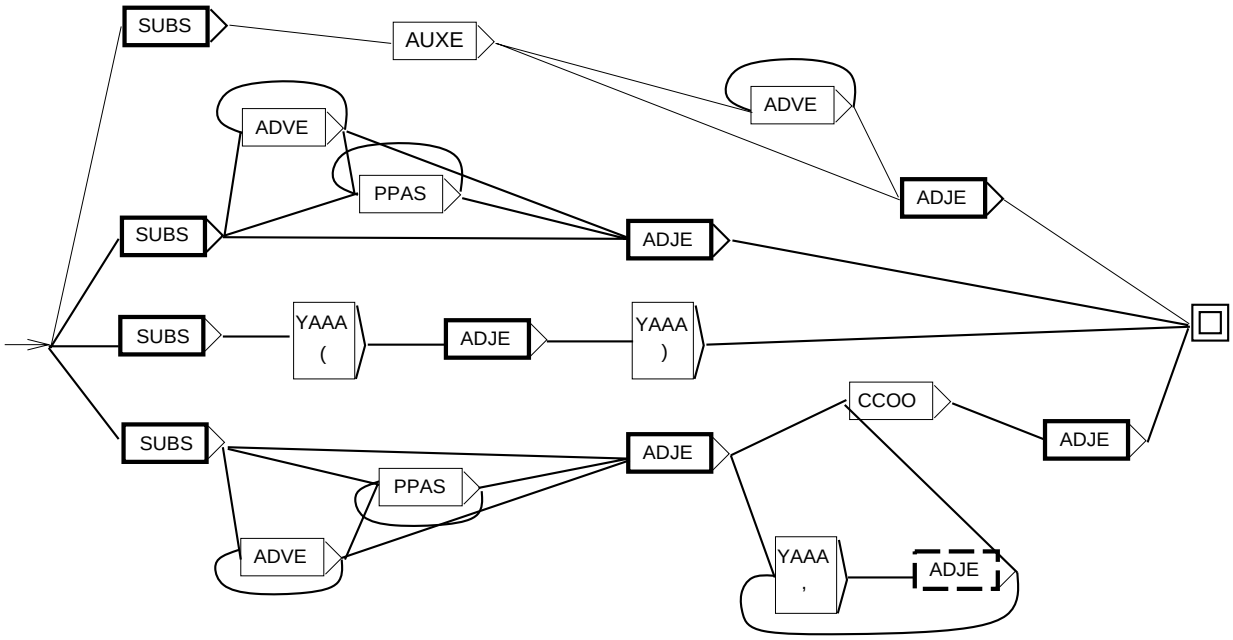

Figure 5: N ADJ

2.3.3 N ADJ finite automaton

In the N ADJ finite automaton, only the lemma of the adjective (`ADJE` tag) is allowed as the second lemma of a pair. The past-participle lemma (`PPAS` tag) is not accepted because pairs of type (N, PPAS) would be mainly incorrect. This decision is motivated by the behavior of the tagger with regard to past-participles: the past-participles employed as adjectives and not followed by a

preposition receive the ADJE tag, except for a few exceptions; those which are followed by a preposition received the PPAS tag. For example, *rayonné* (*radiated*) receives the PPAS tag in the sequence *puissance rayonnée et ...* (*radiated power and ...*) and receives the ADJE in the sequence *champ rayonné en espace libre.* (*free space radiated field*). Also, we did not take into account left coordinations such as *services et systèmes spatiaux spatial services and systems*), since those are ambiguous in most cases. We accept the possibility of encountering the adjective in an attribute position: this transformation does not characterize most frozen compounds, or linguistic co-occurrences, but is accepted by terms: for example, the term *antenne parabolique parabolic antenna*) accepts the transformation: *cette antenne est parabolique* (*this antenna is parabolic*). The count of these co-occurrences is considered only for monolingual extraction; for bilingual extraction purposes, these co-occurrences must be ignored.

2.4 Program description

Our program uses finite automata both to plot the frequencies of terms co-occurrences and to extract them. All the morphosyntactic sequences of a pair that have been encountered in the corpus for a given pattern are entered under the relevant pattern heading with their number of occurrences. In addition, for each occurrence of a pair, its position in the corpus is indicated: this information allows a quick return to the corpus, and can be used for bilingual statistics. Sample output of the program is reproduced bellow:

```
no 65 charge trafic nbc= 6  dist= 3.166667  mdist= 2.166667  var= 0.138889
occ1= charge de trafic patt1= SUBSFS PDEA SUBSMS nbo1= 5
Pos1=  5865/4/7/9 - 5865/29/16/18 - 5965/21/34/36 - 6243/14/19/21 - 6243/16/18/20
occ2= charge normale de trafic patt2= SUBSFS ADJEFS PREPMS SUBSMS nbo2= 1
Pos2=  5965/22/67/70

no 70 bande garde nbc= 6  dist= 3.000000  mdist= 2.000000  var=0.000000
occ1= bandes de garde patt1= SUBSFP PDEA SUBSFS nbo1= 3
Pos1=5872/10/7/9 - 5918/15/31/33 - 6033/9/52/54 -
occ2= bande de garde patt2= SUBSFS PDEA SUBSFS nbo2= 3
Pos2=  5889/6/10/12 - 5889/7/8/10 - 6239/6/42/44
```

This output may be interpreted as follows:

- `no 65`: integer identifier of the pair ,
- *charge* (*loading*) and *trafic* (*traffic*) are the two lemmas of the pair,
- `nbc= 6`: number of occurrences of the pair,
- `dist= 3.166667  mdist= 2.166667  var= 0.138889`: these numerals are the distance measures computed for each pair (see 3.4 and 5.4),
- the pair has been encountered under two patterns (`patt1`, `patt2`) associated with two text sequences (`occ1`, `occ2`). The number of occurrences of each sequence is specified (`nbo1`, `nbo2`). The positions are expressed in the following format: file identifier, sentence identifier, position of the first and last item of the extracted sequence.

The program uses binary trees associated with a hash table to build the pair list and the miscellaneous information that characterize them. The program runs very quickly: it took 2 minutes to extract the 8 000 candidate pairs from the STH corpus for the N_1 *de* (DET N_2 pattern on a Sparc ELC (SS1) station under Sun-Os 4.1.3. release. The program is written in C norm ANSI POSIX.

2.5 Program results

We have plotted the co-occurrences of our two corpora STH (200 000 words) and LBC (800 000 words) and for our two patterns, N_1 (PREP (DET)) N_2 (the parenthesis indicates the optionality of one or several syntactic tags) and N ADJ. An occurrence of a pair is a co-occurrence where the two items of the pair fit with one of these patterns. The tables below summarize the co-occurrence frequencies expressed in pair numbers; so, for the STH corpus and the N ADJ pattern, we have plot 4 483 pairs of which 3 144 have only one occurrence, 655 two occurrences and 684 more than two.

STH	1 occurrence	2 occurrences	¿ 2 occurrences	Total
N ADJ	3 144	655	684	4 483
N_1 (PREP (Det)) N_2	6 834	1 503	1 616	9 953

LBC	1 occurrence	2 occurrences	¿ 2 occurrences	Total
N ADJ	5 201	1 507	2 113	8 821
N_1 (PREP (Det)) N_2	12 167	3 481	6 288	21 936

The extracted co-occurrences for the N_1 (PREP (Det)) N_2 pattern give no information about the representativeness of the different elementary types: N_1 *de* (DET) N_2, N_1 *à* (DET) N_2, N_1 PREP N_2 (with PREP $\neq$ *de* and *à*) and N_1 N_2. Thus, we have tabulated the co-occurrences of each of these elementary types. The results in terms of numbers of pairs are summarized in the following tables:

STH	1 occurrence	2 occurrences	¿ 2 occurrences	Total
N_1 DE[a] N_2	5 393	1 195	1 374	7 962
N_1 À[b] N_2	1 156	161	115	1 432
N_1 Prep[c] N_2	891	132	116	1 139
N_1 N_2	309	66	60	435
Total	7 749	1 554	1 665	10 968

LBC	1 occurrence	2 occurrences	¿ 2 occurrences	Total
N_1 DE[a] N_2	9 558	2 804	5 308	17 670
N_1 À[b] N_2	2 358	492	471	3 321
N_1 Prep[c] N_2	1 474	341	439	2 254
N_1 N_2	682	180	352	1 214
Total	14 072	3 817	6 570	24 459

[a] DE = {*de, d', du, des, de la*}
[b] À = {*à, au, aux, à la*}
[c] Prep ≠ {DE, À}

The results obtained for the N_1 (PREP (DET)) N_2 sub-patterns show the following:

- the pairs of pattern N_1 *de* (DET) N_2 are by far the most numerous,
- the pairs of pattern N_1 *à* (DET) N_2 are as numerous as the pairs of pattern N_1 PREP N_2 (with PREP ≠ *de* and *à*),
- the pair of pattern N_1 N_2 are the least numerous.

Alternatively, if we add the pair numbers corresponding to the elementary types, we obtain a number of pairs superior to the number of pairs plotted with the general pattern N_1 (PREP (DET)) N_2. This is, because numerous pairs share the same lemmas and only vary with the presence or otherwise of a preposition or with the form of the preposition. For example, for the pair (circuit, hyperfréquence) which has the general pattern N_1 (PREP (DET)) N_2, we find co-occurrences belonging to different elementary types: *circuit hyperfréquence* (*microwave circuit*) belonging to the N_1 N_2 elementary type and *circuit à hyperfréquences* of the N_1 *à* (DET) N_2 elementary type. This general pattern has the advantage of producing a census of morphosyntactic variants under a pair. For each pair, we also observe the following:

- co-occurrences which do not refer to the same concept: under the pair (centre, origine), we find the co-occurrences *centre d'origine* and *centre à l'origine*. In the former, *à l'origine* is a sub-sequence of the composed preposition *à l'origine de* and the extracted co-occurrence does not refer to any concept and thus, does not refer to the concept evoked by *centre d'origine*,
- co-occurrences belonging to different elementary types and for which it is difficult, for a monolingual point of view, to decide whether they refer or not to the same concept. Two examples are given below:
  - under the pair (liaison, satellite), the co-occurrences *liaison par satellite* (*satellite link*) belonging to the N_1 *par* N_2 elementary type, and *liaison à satellites* belonging to the N_1 *à* (DET) N_2 type. Do these refer to the same concept ?
  - under the pair (terminal, interface), we pose the same question for the co-occurrences *terminal d'interface* and co-occurrences *terminal à interface*.

A decision about the conceptual status of the different co-occurrences should be made by examination of the original text.

This general problem arises even for co-occurrences belonging to the same elementary type: the pair (couleur, fond) with the N_1 *de* (DET) N_2 pattern groups together *couleur de fond* and *couleur du fond* which do not seem to refer to the same entity. We must answer the following question: should we plot the co-occurrences with a general pattern as N_1 (PREP (DET)) N_2 one and collect some co-occurrences that refer to different entities, or must we plot co-occurrences with a precise elementary type, thus losing links between the different elementary types and, eventually to face the same problem but at a smaller scale. This problem of counting significance is inherent to our quantitative approach and we prefer to have more representativeness, although this may mean recording under a pair a number of occurrences referring to different entity, than to limit the counts and to lose number of essential informations.
If it is important to be aware of the counting significance which has be exposed above, the problem of the missed relevant co-occurrences is also crucial. We have already seen that, in order not to discord the occurrence counts, the binary terms which are over-composed or post-modified are not taken into account. This position implies the non-recognition of some co-occurrences: for the sequence *ondes à polarisation rectiligne perpendiculaires* (*orthogonally polarized waves*), the occurrences of the pairs (onde, polarisation) for the pattern N_1 (PREP (DET)) N_2 and (polarisation, rectiligne) for the pattern N ADJ are plotted, but the occurrence of the pair (onde, perpendiculaire) is ignored. Nevertheless, the adjective *perpendiculaires* modifies *ondes à polarisation rectiligne* and not the simple noun *onde* and not plotting this occurrence is rather a good thing. However, other occurrences belonging to our patterns of terms are not recognized such as the following:

- some occurrences are separated by disruptive items as for example the sequence *une composante (x) brouilleuse*; the presence of an item between parenthesis which appears between *composante* and *brouilleuse* does not allow one to recognize this occurrence of the pair (composante, brouilleuse),

- the tagger makes certain errors that results in the non-recognition of some morphosyntactic sequences; for example, in the sequence *les polarisations quasi circulaires*, the occurrence of the pair (polarisation, circulaire) is not recognized. The automaton asks for agreement in gender and number between the noun and the adjective that modifies it: in this sequence *polarisations* receives a correct tag (plural feminine noun), but *circulaires* receives a incorrect one, namely plural masculine adjective. As agreement in gender that is required by the automaton is not present, this occurrence is not recorded.

The examples where the co-occurrences are not plotted show that:

1. It is impossible to take into account all the morphosyntactic sequences where correct co-occurrences appear without taking into account incorrect co-occurrences,

2. The selection of co-occurrences relies upon the tagging; tagging errors can imply the non-recognition of good co-occurrences as the selection of bad ones.

We will evaluate the statistical scores presented in the following section only upon pairs that accept two occurrences. That frequency threshold is the same as the one given by [Lafon, 1984] but it is low compare to the threshold of five which is recommended by [Smadja and McKeown, 1990] or [Church and Hanks, 1990]. We take into account the preliminary linguistic filtering: "noise" significantly decreases. This threshold is arbitrarily fixed; but, the values of the statistical scores are not relevant for pairs which accept only one occurrence. By throwing away pairs with one occurrence, we retain only half of the pairs. We are aware that some of the co-occurrences could be terms and thus the thresholding process can result in the loss of desired information.

Now, we reach the statistical part of our work. We have described how statistical scores should be guided: we have filtered out co-occurrences that are all possible terms from a morphosyntactic point of view. First, we are going to present a few statistical scores (section 3). In section 4, these scores will be evaluated by a graphical method for a sub-set of our pairs: those of N_1 *de* (DET) N_2 pattern extracted from the STH corpus. We will then take into account the results of this evaluation and we will examine the conceptual sort, for our two general patterns: N_1 (PREP (DET)) N_2 and N ADJ, which is proposed by these scores (section 5). We will also examine, in section 5, the results of various statistical scores that do not analyze the strength of the bond between the two lemmas of a pair, but rather provide other types of information.

3 Statistical scores

Some of the statistical scores that have been used are well-known. We recall here their definitions.
Four types of numeric characteristics are computed :

- frequencies,
- association criteria,
- Shannon diversity,
- distance measures.

To these numeric characteristics, we can adjoin a measure which uses bilingual data: affinity. We are not going to describe it; it is studied extensively in [Gaussier, 1994].
The previous numeric characteristics do not behave identically: frequencies are the parameters of the association criteria, association criteria and affinity compute the strength of the bond of the two lemmas of a pair and propose a conceptual sort of them; diversity and distance measures are not discriminatory scores but provide other types of informations.

3.1 Frequencies

The occurrences that are plotted for a given morphosyntactic structure are:

- the number of occurrences of a pair; this number has been computed by the program that count and extract term co-occurrences. The following frequencies are computed with the set of pairs extracted:
- the number of occurrences of pairs where a given lemma appears as the first element of the pair,
- the number of occurrences of pairs where a given lemma appears as the second element of the pair,
- the total number of occurrences of the pairs (for each syntactic pattern).

These four numbers are used by all the statistical scores that we have chosen.

3.2 Association criteria

From a statistical point of view, the two lemmas of a pair could be considered as two qualitative variables whose link has to be tested. A contingency table is defined for each pair (L_i, L_j):

	L_j	$L_{j'}$ with $j' \neq j$
L_i	a	b
$L_{i'}$ with $i' \neq i$	c	d

where:

a stands for the frequency of pairs involving both L_i and L_j,

b stands for the frequency of pairs involving L_i and $L_{j'}$,

c stands for the frequency of pairs involving $L_{i'}$ and L_j,

d stands for the frequency of pairs involving $L_{i'}$ and $L_{j'}$.

The sum $a+b+c+d$, written N, is the total number of occurrences of all pairs obtained for a given pattern.

The statistical literature proposes many scores which can be used to test the strength of the bond between the two variables of a contingency table. These scores are either link scores (SMC), or statistical tests (Φ^2) that we used as link scores. We do not make the difference between link scores and statistical tests. Furthermore, we use these latter, no to state upon the dependency or the independence of two qualitative variable, but only as link score: the two lemmas of a pair belong to a morphosyntactic structure and are not independent. In our hypothesis, these statistical scores are expressed thanks to the value of a contingency table. We do not examine if the values taken by the scores are statistically relevant: none of the pairs is rejected. We are going to enumerate the statistical scores that we have tested in our experiments:

- some have already been used in lexical statistics: simple matching coefficient (equation 1), Φ^2 coefficient (equation 7), log-likehood coefficient (equation 10), association ratio (concept of mutual information) (equation 8),
- some have been used in other technical domain as for example biology: Kulczinsky coefficient (equation 2), Ochiai coefficient (equation 3), Fager and McGowan coefficient (equation 4), Yule coefficient (equation 5), McConnoughy coefficient (equation 6),
- we have introduced a new score for reason that we will explain: cubic association ratio (equation 9).

Simple Matching Coefficient (SMC)

This score is symmetrical with L_1 and L_2, and varies from 0 to 1.

$$SMC = \frac{a+d}{a+b+c+d} \quad (1)$$

Kulczinsky Coefficient (KUC)

This score varies from 0 to 1. When L_1 (resp. L_2) is only observed with L_2 (resp. L_1), the value of KUC exceeds 0.5.

$$KUC = \frac{a}{2}(\frac{1}{a+b}+\frac{1}{a+c}) \quad (2)$$

Ochiai Coefficient (OCH)

This score varies from 0 to 1.

$$OCH = \frac{a}{\sqrt{(a+b)(a+c)}} \quad (3)$$

Fager and McGowan Coefficient(FAG)

This score varies from a non definite value that *a priori* could be negative to 1 (this superior born is never reached).

$$FAG = \frac{a}{\sqrt{(a+b)(a+c)}} - \frac{1}{2\sqrt{a+b}} \quad (4)$$

Yule Coefficient (YUL)

This score varies from -1 to +1. It equals +1 when one of the lemmas appears always with the other one.

$$YUL = \frac{ad-bc}{ad+bc} \quad (5)$$

McConnoughy Coefficient (MCC)

This score varies from -1 to +1

$$MCC = \frac{a^2 - bc}{(a+b)(a+c)} \quad (6)$$

It can be observed that:

$$MCC = 2KUL - 1$$

and deduced that McConnoughy coefficient and Kulczinsky one accept a similar distribution. For this reason, we do not proceed at the evaluation of this score.

Φ^2 Coefficient (PHI)

This score has been used by [Gale et Church, 1991b] to align words inside align sentences.

$$\Phi^2 = \frac{(ad - bc)^2}{(a+b)(a+c)(b+c)(b+d)} \quad (7)$$

Association ratio

This method of measuring the association ratio of two words has been described by [Brown *et al.*, 1988] for the extraction of bilingual resources and by [Church and Hanks, 1990] for monolingual extraction. The association ratio of a lemma pair (L_1, L_2) is based upon the concept of mutual information. This score, noted IM, compare the probability to observe two lemmas together with the probability to observe the two lemmas alone. Its definition is:

$$IM(L_1, L_2) = \log_2 \frac{P(L_1, L_2)}{P(L_1)P(L_2)}$$

where P is the probability.
This definition is equivalent to the following expressed with our values:

$$IM = \log_2 \frac{a}{(a+b)(a+c)} \quad (8)$$

Association ratio as it is defined give to much pound to rare events (see section 5.2.1).Thus, we have introduced the following score:

Cubic association ratio (IM^3)

This formula results from an experimental study: we have decided to give more pound to frequent events and we have tried all the power of a from 2 to 10. Cube has been retained because it appears as a good compromise between the fact to retain only rare events and to neglect them to much.

$$IM^3 = \log_2 \frac{a^3}{(a+b)(a+c)} \quad (9)$$

Log-Likehood Coefficient

This coefficient introduced by [Dunning, 1993] is the test of the log-likehood ratio applied to a binomial law. It is equivalent in our hypothesis to a generalized mutual information and thus, could be expressed thanks to our values of contingency table:

$$\begin{aligned} Log-Likehood &= a\log a + b\log b + c\log c + d\log d \\ &\quad -(a+b)\log(a+b) - (a+c)\log(a+c) \\ &\quad -(b+d)\log(b+d) - (c+d)\log(c+d) \\ &\quad +(a+b+c+d)\log(a+b+c+d) \end{aligned} \tag{10}$$

A property of these scores is that their values increase with the strength of the bond of the lemmas.

3.3 Shannon diversity

This score has been introduced by [Shannon, 1948] and is used in biology to classify entities into species. It seems interesting to us because it allows to compare the role of a lemma in a frozen position inside a pair with the set of pairs. The idea being that a lemma which appears in an important number of pairs in equal proportion in the first position of a pair is either a item which always allow to create a term, as for example the noun *système* (*system*), or the contrary, saying a noun that never leads to the creation of a new term as for example the noun *caractéristique* (*characteristic*). Diversity characterizes the marginal distribution of the elements of a couple through the range of couple. To compute it, you need a contingency table of $n \times m$ dimensions whose theoretical representation is:

x_i y_j	y_1	y_2	..	y_j	..	y_m	Total
x_1	n_{11}	n_{12}	..	n_{1j}	..	n_{1m}	$n_{1.}$
x_2	n_{21}	n_{22}	..	n_{2j}	..	n_{2m}	$n_{2.}$
.	.	.	.	.	.	.	.
x_i	n_{i1}	n_{i2}	..	n_{ij}	..	n_{im}	$n_{i.}$
.	.	.	.	.	.	.	.
x_k	n_{k1}	n_{k2}	..	n_{kj}	..	n_{km}	$n_{k.}$
Total	$n_{.1}$	$n_{.2}$	..	$n_{.j}$	..	$n_{.m}$	$n_{..}$

with X and Y characters, x_1, x_2, ..., x_i, ..., x_k, k modalities of X; y_1, y_2, ..., y_i, ..., y_m, m modalities of Y. The n observations are distributed following the modalities of X and Y. In our application, X represents the syntactic tag of the first lemma (N) and Y the syntactic tag of the second lemma of a pair (N or ADJ). The number n_{ij} which appear at the intersection of the line i and the column j of the table is the number of occurrences of the pairs with the lemma x_i with the tag X and the lemma y_j with the tag Y. This is a part of the contingency table associated to the N ADJ pattern:

N_i Adj_j	*progressif*	*circulaire*	*porteur*	...	Total
onde	19	4	6	...	$nb_{(onde,.)}$
limiteur	9	0	0	...	$nb_{(limiteur,.)}$
cornet	0	2	0	...	$nb_{(cornet,.)}$
...	...	...	...	...	...
Total	$nb_{(.,progressif)}$	$nb_{(.,circulaire)}$	$nb_{(.,porteur)}$	...	$nb_{(.,.)}$

The line counts $nb_{i.}$, which are found in the last column, represent the distribution of the adjectives with regards to a given noun. The columns counts $nb_{.j}$, which are found on the last line, represent the distribution of the nouns with regards to a given adjective. These distributions are called "marginal distributions" of the nouns and the adjectives for the N ADJ structure. Diversity is computed for each lemma appearing in a pair, using the formula:

$$\begin{aligned} H_i &= nb_{i.} \log n_{i.} - \sum_{j=1}^{s} nb_{ij} \log nb_{ij} \\ H_j &= nb_{.j} \log n_{.j} - \sum_{i=1}^{s} nb_{ij} \log nb_{ij} \end{aligned} \tag{11}$$

For example, using the contingency table of the N ADJ structure above, diversity of the noun *onde* (*wave*) is equal to:

$$\begin{aligned} H_{(onde,.)} = nb_{(onde,.)} \log nb_{(onde,.)} - (&nb_{(onde,progressif)} \log nb_{(onde,progressif)} + \\ &nb_{(onde,porteur)} \log nb_{(onde,porteur)} + \ldots) \end{aligned}$$

We note H_1, diversity of the first lemma of a pair and H_2 diversity of the second lemma. We take into account the diversity normalized by the number of occurrences of the pairs:

$$\begin{aligned} h_i &= \frac{Hi}{n_{ij}} \\ h_j &= \frac{Hj}{n_{ij}} \end{aligned} \tag{12}$$

The normalized diversities h_1 and h_2 are defined from H_1 and H_2. We will see later on how we use the normalized diversity to correct some errors done by the tagger.

3.4 Distance Measures

French base-terms often accept modifications of their internal structure as it has been demonstrated previously. Each time, an occurrence of a pair is extracted and counted, two distances are computed: the number of items *Dist* and the number of main items *MDist* which occur between the two lemmas. Then, for each couple, the mean and the variance of the number of items and main items are computed. The variance formula is:

$$V(X) = \frac{1}{n} \sum (x_i - \bar{x})^2$$
$$\sigma(X) = \sqrt{V(X)}$$

3.5 Affinity score

Word alignment inside bilingual aligned sentences give for each word of a language a list of possible associated translations (see [Gaussier *et al.*, 1992]). Affinity score takes advantage of the bilingual associations of words to evaluate the dependency between two words of the same language. The idea beyond is that two words which form a term are likely to share a lot of associations. For example, *pomme* and *terre* in French would probably both contain *potato*, as well as several other words related to the context, such as *crop*, in their lists. To resume, affinity score is a link measure between words which is defined from a link measure between lists.

4 Graphical evaluation of statistics

Each score proposes a conceptual sorting of the pairs. This sorting, however, could put at the top of the list (with the highest score) compounds that belong to general language rather than to technical terminology. Since our goal is to find out which score is the more adequate to extract terminology, it is essential to evaluate the correlation between the score values and the "termhood" of the pairs. Therefore, we compare the values obtained for each score to a reference terminological list of our technical domain. This reference list is described below. This evaluation has been done for 2 200 French pairs[3]of N_1 *de* (DET) N_2 structure extracted from our corpus *STH* (200 000 words). If the pair is found in the reference list, it is considered as a good candidate, else as a bad one. Then, we can examine the goodness of the candidates with respect to the different scores and decide which scores are the best for detection of terminological concepts.

4.1 Reference list

A reference list of the terms of the domain of telecommunications could be built by hand from our corpus *STH* in the same way that [Van der Eijk, 1993] did for his bilingual terminology extraction program. Another possibility consists on obtaining a terminology bank of the domain. We have used Eurodicautom, the terminology data bank of the EEC, telecommunication section, which has been elaborated by experts. This terminology bank is available in the nine languages of the European Community . We obtained a list of over 6 000 terms. This is a sample of this list:

```
caractéristiques fondamentales d'une attribution de fréquence
température apparente du ciel
```

[3]We remind the reader that only pairs which appear at least twice in the corpus have been retained.

```
température cinétique
température de bruit d'un récepteur
température de bruit d'un système de réception
température de bruit de fonctionnement
température de bruit de fond
température de bruit équivalente
température de bruit quantique
température de bruit thermique
température de régime
température équivalente d'antenne
température équivalente pour le trajet descendant
température moyenne de rayonnement
```

This sample shows that all the terms are single line entries: each term occupies a line of a text file. There is no internal architecture. We can remark on this sample that six terms are built with the term *température de bruit* (*noise temperature*), but it does not appear in the list; the remark is true for *température équivalente*, *bruit de fonctionnement*, *trajet descendant*, *température moyenne*. Furthermore, some sequences as *température apparente du ciel* or *caractéristiques fondamentales d'une attribution de fréquence* look like more as free nominal phrase than terms. We compared our candidate pairs to this reference list, deciding that our candidates of N_1 *de* (DET) N_2 structure would be considered "good" if they were a substring of one of the term in the list. This first evaluation gave bad results: only 300 pairs of the 2 200 were considered good candidates. We attributed this fact to defects of the reference list and, decided to complement said list with a list of terms extracted directly from our corpus *STH*. For that purpose, We gave the list of pairs which do not appear in Eurodicautom (1 900 pairs) to three experts of the telecommunication domain. We selected those pairs for which at least two judges agreed on their goodness, another 900 pairs (only 300 pairs obtained unanimity of the judges) which we included in the reference list. To conclude, we assume that 1 200 pairs of our list of 2 200 candidate pairs are terms of the telecommunication domain.

4.2 Graphical evaluation

Each score yields a list where the candidates are sorted according to the decreasing score value. We have divided this list in equivalence classes which generally contain 50 successive pairs (the classes which contain more than 50 pairs are presented below).. The results of a score are represented graphically as an histogram in which the x-axis represents different classes, and the y-axis the ratio of good pairs. If all the pairs in a class belong to the reference list, we obtain the maximum ratio of 1; if none of the pairs appear in the reference list, the minimum ratio of 0 is reached. The ideal score should assign high (resp. low) values to good (resp. bad) pairs, i.e. candidates which belong (resp. don't belong) to the reference list. In other words, the histogram of the ideal score should assign to equivalence classes containing the high values (resp. low values) of the score a ratio close to 1 (resp. 0). Furthermore, an ideal score should allow to determine a threshold or several thresholds under which we can

discard candidates. The histogram of an ideal score should have zero or few classes taking intermediate values. A ideal histogram, with only one threshold, could be the following:

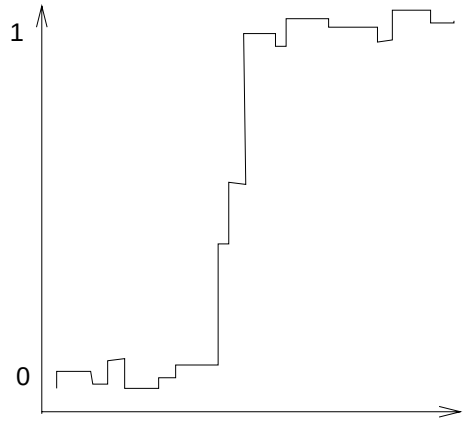

Figure 6: Ideal histogram

For each score, we will therefore smooth the histogram to a curve and examine if this latter looks like the ideal curve:

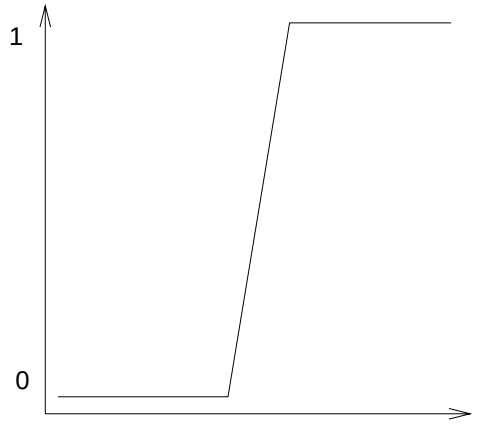

Figure 7: Ideal curve

When an histogram is irregular as the following:

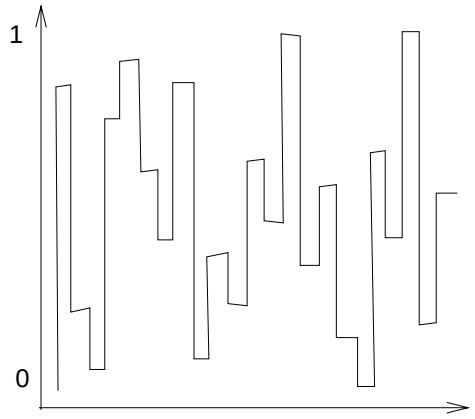

Figure 8: Irregular histogram

it is impossible to deduce a curve and we know that the score does not allow to select good candidates.
Before examining the histograms of the scores we have presented in section 3, we want to precise the two following points:

- we have defined equivalence classes of generally 50 successive items belonging to the list of candidates for a score. Nevertheless, two adjacent equivalence classes must not contain candidates that share an identical value of the score. For example, for the association criteria, the equivalence class containing the highest values, between 13,65 and 11,19 contains 51 candidates as the fiftieth and the fiftieth-first share an identical value. These classes are generally quite homogeneous except for the frequency of a pair: the equivalence class of the pairs with two occurrences contains 1 500 candidates.

- the comparison with the reference list has been done on a subset of the pairs. The evaluation concerns 2 600 pairs and the list of reference has been built on 2 200 pairs. The representativity of an equivalence class never exceeds 0,8 and some scores whose maximal value tend towards 0,6 are retained.

We give now the histograms of the following scores:

- Figure 9

  N1 Number of occurrences of pairs where a given lemma appears as the first element of the pair

  N2 Number of occurrences of pairs where a given lemma appears as the second element of the pair

  NC Number of occurrences of a pair

PHI2 Φ^2 Coefficient

OCH Ochiai Coefficient

YUL Yule Coefficient

- Figure 10

  FAG Fager and MacGowan Coefficient

  AFF Affinity

  h1 Normalized diversity applied to N_1

  h2 Normalized diversity applied to N_2

  H1 Diversity applied to N_1

  H2 Diversity applied to N_2

- Figure 11

  LOG Log-Likehood Coefficient

  IM Association ratio

  IM2 Association ratio (numerator at square)

  IM3 Association ratio (numerator at cube)

  KUC Kulczinsky Coefficient

  SMC Simple Matching Coefficient

At first sight, these histograms all show a general growing trend that confirm that the score values increase with the strength of the bond between the lemma of a pair. However, the growth is more or less clear, with more or less sharp variations. Let us examine in details the histogram shapes:

- For SMC, the curve goes up, then done, and stabilizes around a mean of 0,3. This level means that for such scores, high values yield more wrong candidates than good ones. The peak found in low values can show that this score reject good candidates. This score is not retained.
- For KUC, the growth is clear, but with important fluctuations around the general slope. It is quite difficult in such case to determine a threshold above which the proportion of good candidates is and remains constantly interesting. This score has little discriminating power and is not retained.
- For OCH, the growth is not clear for high values and important fluctuations appear. This score is not retained.
- For FAG, the slope is negative at the beginning and then grows steadily. This curve isolates two areas of values, low ones with a low proportion of good candidates and high ones with a greater proportion. This score can play a discriminatory role and we have retained it.
- For YUL this curve is a plateau with an average value of 0,3. Such a score seems hard to be exploited and we did not retain it.
- For PHI2, the curve shows three parts: steady growth with numerous variations, slow growth and again steadier growth. As was the case for FAG, it is possible to identify an acceptability threshold, but such threshold would retain a limited number of candidates. We did not retain this score.
- For LOG, the curve grows slowly at first and then decidedly. There is a true opposition in the behaviors of small and big values. Moreover, the representativity of good candidates tends towards its maximum (0,8). Yhis score is retained.
- For IM, the curve comes to a plateau around 0,35 with variations. This score is not retained.
- For IM2, the curve is not marked with numerous variations. This score is not retained.
- For IM3, the growth is clear and we obtain a good opposition between low and high values. We have retained this score.
- The histogram of NC (the number of occurrences of the pair) comes closest to the ideal curve: it shows a clear and sharp growth towards maximal values. However, the number of items of a class is not constant and a close inspection reveals that a certain number of good candidates creep in the class with lowest value. This score is not ideal since it does not

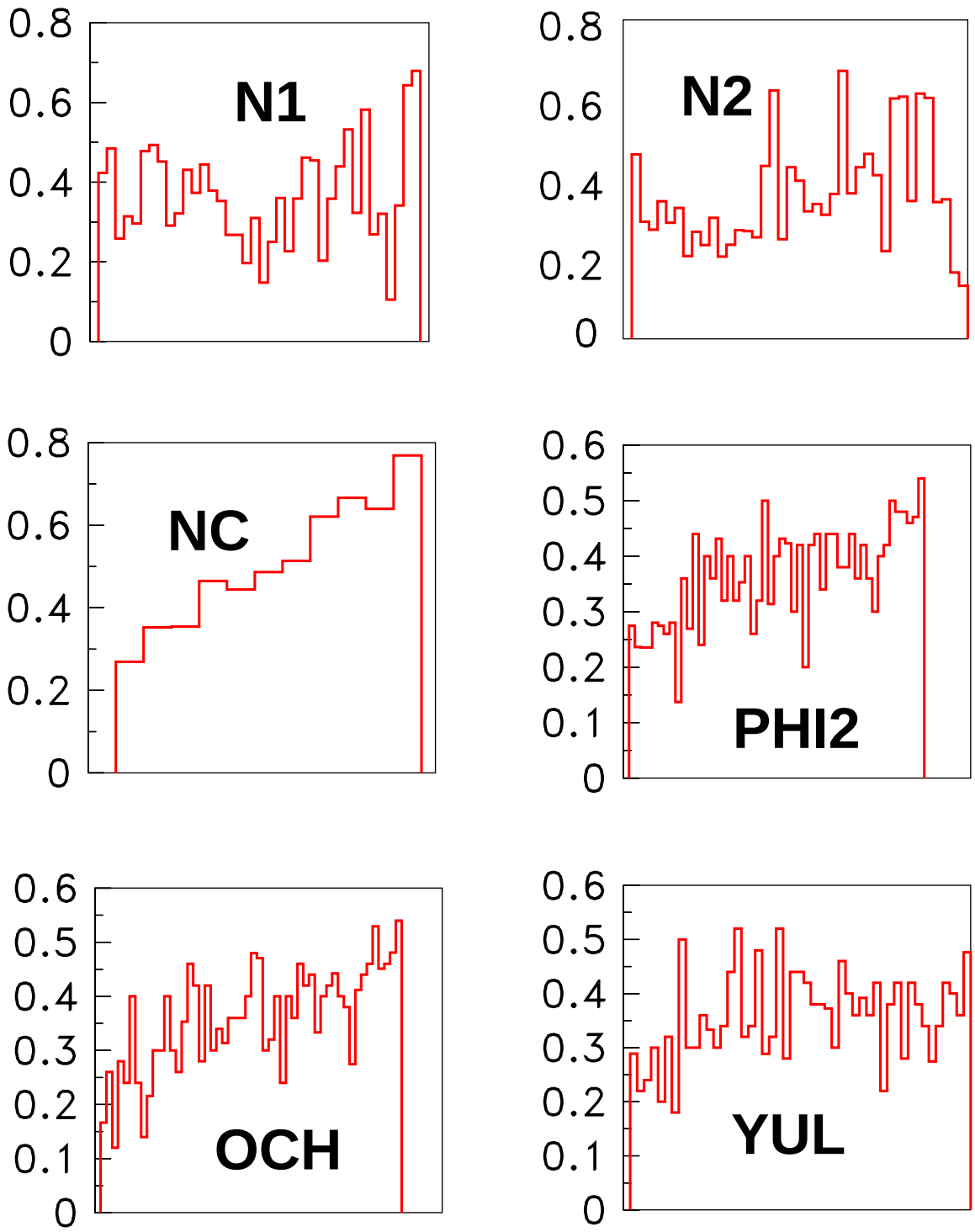

Figure 9: Histograms of scores (part 1/3)

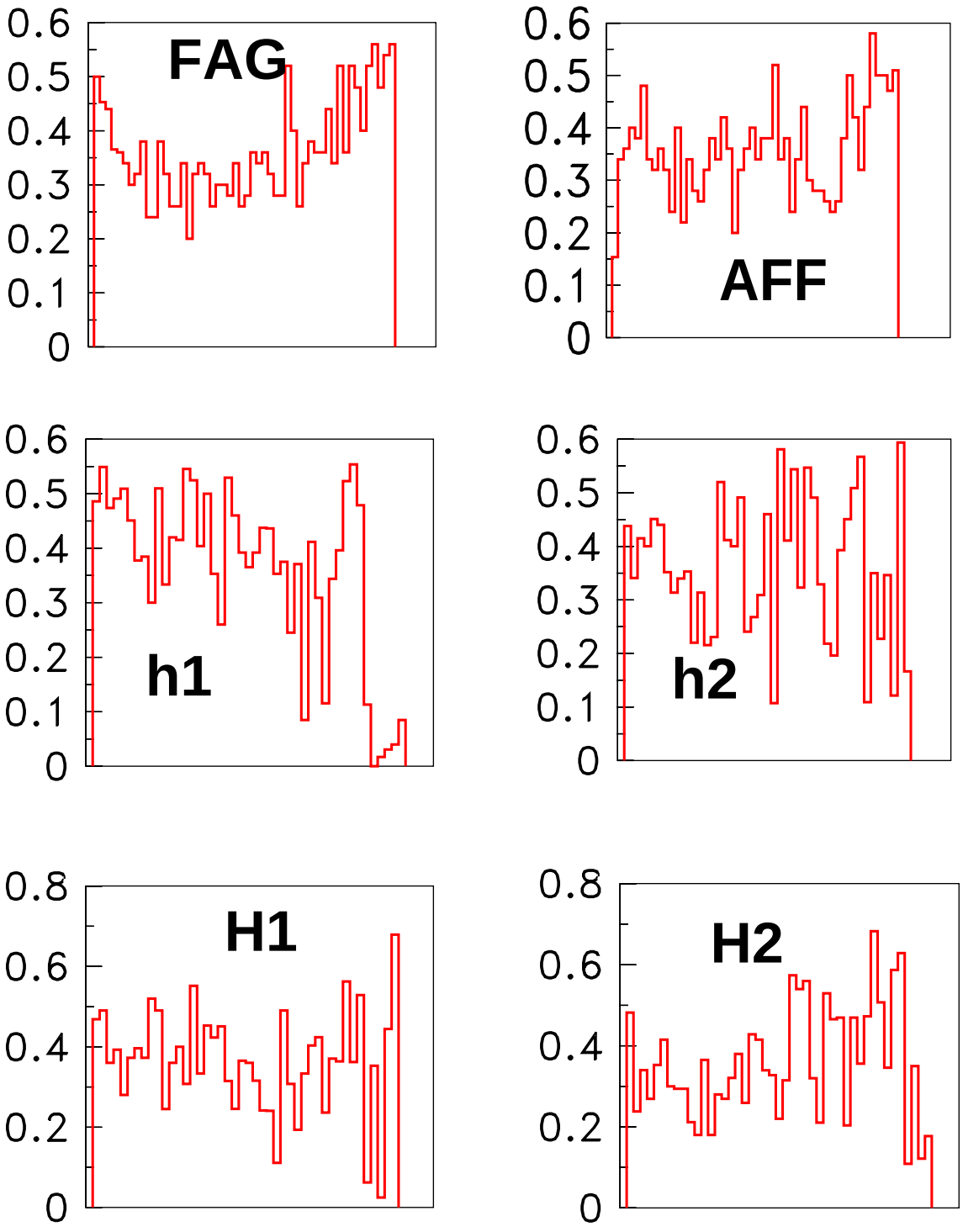

Figure 10: Hostograms of scores (part 2/3)

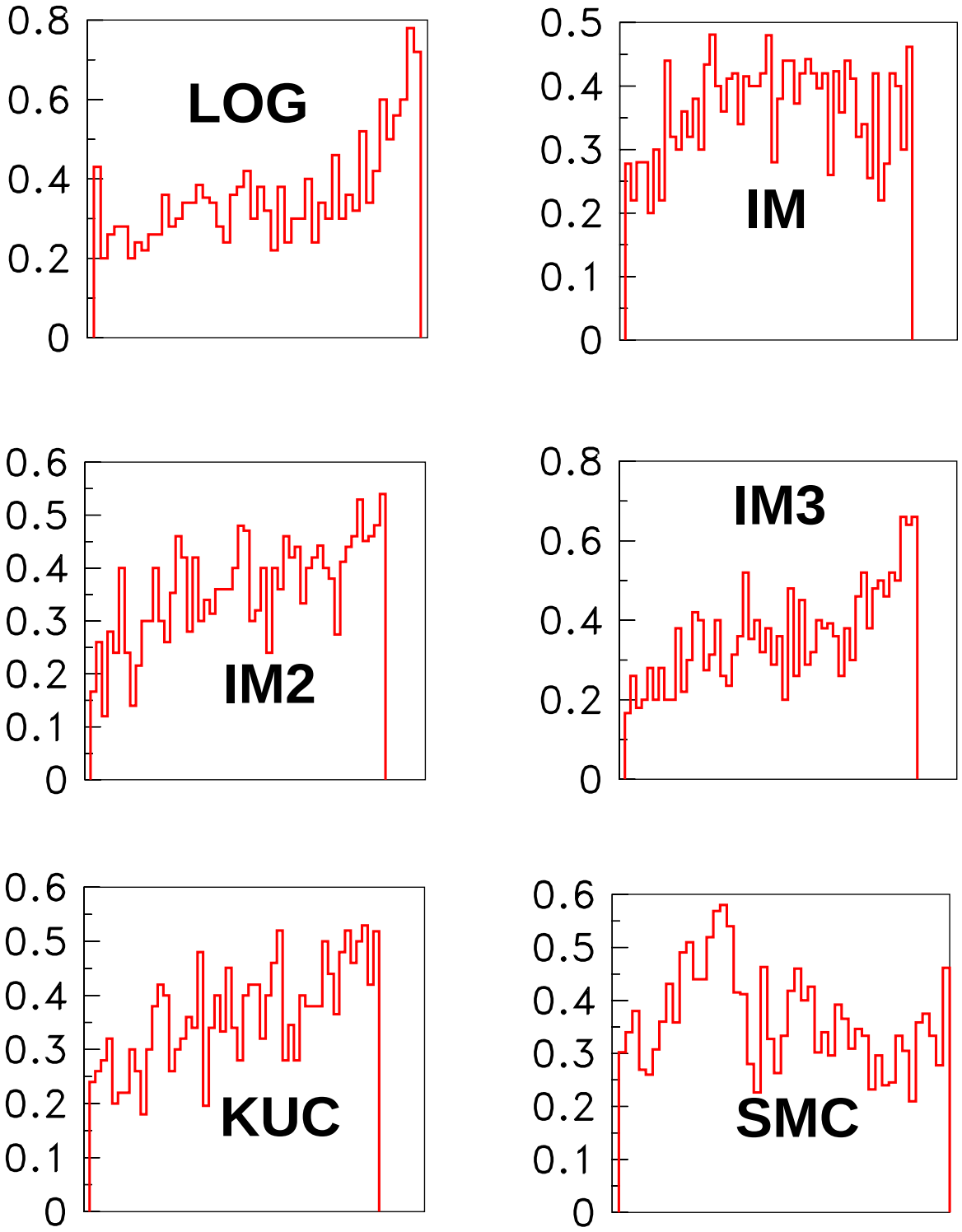

Figure 11: Histograms of scores (part 3/3)

make the whole set of good candidates stand out, but we have obviously retained it.

- The histograms of N1 and N2 show no clear growth: the fact that a lemma appears often either in first or in second position of a pair is not discriminatory.

- For AFF, the curve tends to show two plateaus, one for low values around 0,3 and the other for high values around 0,45. This score is not retained.

- For h1, h2, H1, H2
  These histograms show no clear growth. The histogram of h1 shows that high values characterize bad candidates. This score is not retained but could be used eventually for filtering.

The outcome of the previous study is that few scoring methods will meet our objectives. The most effective in a view to extract terminology seem to be the log-likelihood coefficient, Fager and MacGowan coefficient, cubic association ratio, and NC, the simple frequency of the pair. If it were for the sole histograms, one could be tempted to keep only the frequency of the pair (NC), but we have remarked that due to its definition, infrequent terms will not stand out. Thus, the problem arises of knowing if these 4 scores are complementary, or on the contrary closely related, i.e. do they isolate the same candidates in the set of pairs ? We have studied the scores through the use of a covariance matrix in order to reveal links of opposition. The results of this study shows that the scores have little correlation between them, with a maximum of 0.35 between IM3 and FAG. It seems therefore difficult to do more spadework to select a score and to eliminate some scores of the remaining ones. This graphical evaluation shows that none of the retained scores meet our objective: none of the histograms shows a clear threshold under which we can discard candidates. We are going to examine more closely the range of pairs proposed by each of these four scores as well as the results provided by the normalized diversity.

5 Examination of the results

The preceding graphical evaluation only retains four scores among the eighteen tested. The elimination of these scores could be easily explained: we have demonstrated that one of the best criteria of selection of a term from a corpus is its number of occurrences whereas most association criteria, among which the association ratio, give importance to pairs with a low frequency. The retained association criteria have the property to not systematically eliminate frequent pairs, as we are going to see. The statistical scores have been applied on our two sets of pairs: those of N_1 (PREP (DET)) N_2 and N ADJ structures. We keep the two corpora STH (200 000 words) and LBC (800 000 words) in order to evaluate if the size of the data is or is not an important parameter. We will examine the results obtained with frequency, then with the association criteria: association ratio, cubic association ratio, log-likehood coefficient and

Fager and MacGowand coefficient. We will present for each of these scores the pairs that receive the highest values. These observations lead us to retain only the log-likehood coefficient. This section will end by the examination of the informations provided by the normalized diversity and by the distance measures. These last two measures are not discriminatory measures but are used together with the association criterion retained.

5.1 Frequency

Frequency appears as one of the best scores to detect the terms of a technical domain. This result contradicts the now widespread opinion that mutual information (association ratio) is a better estimate of word association than simple frequency counts (see for example [Church et Hanks, 1989]). All the frequent pairs that we obtained yield a low value of association ratio, whereas there is no doubt about their terminological nature. This bad behavior of the association ratio is an important result and is commented below (see 5.2.1). If the frequency results contradict several works in lexical statistics, they confirm the work of [Enguehard, 1992] on automatic natural acquisition who takes only into account frequency of words in corpora. Moreover, if the important number of occurrences of a pair is a determinant point, the important number of pairs where a given lemma appears in first position (resp. in second position) is not discriminatory at all (cf. histograms N1 and N2 - figure 9). These results could contradict those of [Bourigault, 1994] who considers that the productivity of the head noun is a discriminatory point to state about the terminological character of a nominal phrase; however, as the author does not only extract base-terms is is true that some nouns, as in our domain the noun *système* (*system*), systematically produce terms of length 2 and more: *système d'alimentation* (*supply system*), *système à débit binaire* (*bit rate system*), *système de télécommunications par satellite* (*satellite communications system*), etc.. However, this behavior is not shared by all the head nouns: *caractéristique* (*characteristic*), *fonctionnement* (*operation, description, ...*), *utilisation* (*use*), *information* (*information*) , ..., do not produce or seldom terms.
The sorting provided by frequency integers very quickly pairs that are not terms: for example, the first bad candidate is the pair (cas, transmission) which appears at the 56th position for the STH corpus. To give an element of comparison, this pair is also the first bad candidate with the sorting provided by Log-Likehood coefficient but it only appears at the 176th position. We will see how some of these wrong pairs could be eliminated by using the informations provided by diversity.
We are going to present the pairs which share the highest values of number of occurrences.

N_1 (PREP (DET)) N_2

The number of occurrences of the pairs take place between 2 and 223 for STH and between 2 and 1 188 for LBC. For each pair, we precise for sake of clarity

the morphosyntactic sequence the most numerous with its frequency indicated inside parenthesis; we precise here neither the variations of prepositions nor the modifiers that could appear with the pair when they are underestimated. (the morphosyntactic sequences which appear under a pair are commented with the distance measures (section 5.4)). Frequency goes down very quickly: for the STH corpus, only one pair owns more than 200 occurrences, 4 pairs own between 100 and 200 occurrences; for the LBC corpus, only one pair owns more than 1 000 occurrences, 3 between 500 and 1 000, 14 between 200 and 500,...The pairs with the highest frequency are given above. We use the following notations:

Nbc Number of occurrences of the pair

IM Association ratio associated to the pair

Corpus	Pair of N_1 (PREP (DET)) N_2 structure	The most frequent pair sequence	**Nbc**	IM
STH	(largeur, bande)	*largeur de bande* (197)	**223**	5,73
	(bande, base)	*bande de base* (142)	**145**	5,52
	(amplificateur, puissance)	*amplificateur(s) de puissance* (137)	**137**	5,66
	(température, bruit)	*température de bruit* (110)	**126**	6,18
	(système, satellite)	*système(s) à satellites* (89)	**108**	1,81
	(télécommunication, satellite)	*télécommunication(s) par satellite* (88)	**99**	4,09
	(réseau, satellite)	*réseau(x) à satellites* (62)	**97**	2,58
	(temps, propagation)	*temps de propagation* (93)	**94**	6,89
	(bande, fréquence)	*bande(s) de fréquences* (85)	**89**	3,54
	(système, signalisation)	*système(s) de signalisation* (82)	**85**	2,81
	(liaison, satellite)	*liaison(s) par satellites* (73)	**82**	2,72

Corpus	Pair of N_1 (PREP (DET)) N_2 structure	The most frequent pair sequence	**Nbc**	IM
LBC	(canal, sémaphore)	*canal/canaux sémaphores* (1 188)	**1 188**	4,77
	(système, signalisation)	*système(s) de signalisation* (839)	**847**	3,60
	(point, sémaphore)	*point(s) sémaphore(s)* (677)	**679**	3,57
	(accusé, réception)	*accusé(s) de réception* (592)	**592**	6,37
	(signal, fin)	*signal/signaux de fin* (385)	**391**	4,20
	(trame, sémaphore)	*trame(s) sémaphore(s)* (354)	**354**	4,55
	(message, adresse)	*message(s) d'adresse* (187) *message initial d'adresse* (120)	**337**	4,14
	(réception, signal)	*réception du signal* (143) *réception d'un signal* (126)	**332**	3,45
	(unité, signalisation)	*unité(s) de signalisation* (320)	**320**	3,63
	(réseau, sémaphore)	*réseau(x) sémaphore(s)* (283)	**283**	3,27
	(connexion, sémaphore)	*connexion(s) sémaphore(s)* (274)	**274**	3,16

N ADJ

The number of occurrences of the pairs take place between 2 and 750 for STH and 2 and 340 for LBC. The pairs of N ADJ structure the most frequent and listed below are not all base-terms of length 2: in the STH corpus, *service fixe* (*fixed service*) is a sub-sequence of the base-term of length 3: *service fixe par satellite* (*fixed-satellite service*). We can remark, here again, that the association ratio takes middle values and do not consider that frequent pairs are terms

of the domains. Nevertheless, these values are more regular for the N ADJ structure than for the N_1 (PREP (DET)) N_2 ones.

Corpus	Pair of N ADJ structure	**Nbc**	IM	Corpus	Pair of N ADJ structure	**Nbc**	IM
STH	(station, terrien)	**750**	3,37	LBC	(service, supplémentaire)	**340**	4,33
	(débit, binaire)	**134**	5,32		(centre, international)	**325**	3,63
	(voie, téléphonique)	**118**	4,75		(équipement, terminal)	**275**	5,43
	(accès, multiple)	**105**	5,66		(considération, général)	**256**	5,38
	(liaison, montant)	**88**	5,17		(circuit, international)	**213**	2,23
	(secteur, spatial)	**79**	4,80		(réseau, national)	**208**	3,26
	(liaison, descendant)	**77**	5,22		(niveau, relatif)	**202**	4,16
	(service, fixe)	**66**	5,33		(entité, fonctionnel)	**199**	5,42
	(engin, spatial)	**57**	5,28		(caractère, graphique)	**196**	5,86
	(station, spatial)	**56**	1,17		(adresse, complète)	**183**	5,30
	(station, distant)	**44**	2,88		(effet, local)	**169**	5,43

5.2 Association criteria

From the ten scores presented in section 3.2, the graphical evaluation allows to retain only three of them: cubic association ratio (formula 9), log-likehood coefficient (formula 10) and Fager and MacGowan coefficient (formula 4). These coefficient have been retained because on the contrary of the other ones, they do not eliminate frequent pairs. Each of the score propose a conceptual sorting of the pairs. We are going to examine these three scores and to try to determine what are the specificities of their sortings.

5.2.1 Associatio ratio and Cubic association ratio

The bas results obtained with the association ratio surprised us: the pairs which own high values of this score share low frequencies (2 to 3). Association ratio isolates frozen compounds as *aiguille d'une montre* (*clockwise*), *béton armé* (*reinforced concrete*), frozen adverbs as *dos à dos* (*back-to-back*) but not terms of the domain. The high value comes from the fact that the two lemmas appear only together and never inside other pairs. The introduction of a strong constraint (we have only kept pairs that followed a specific linguistic pattern and with a frequency of more than 2) has probably changed the results with these statistics that were until now performed on raw corpora, without linguistic annotations. These bad results of the association ratio lead us to empirically modify the formula : we put the numerator at the power of 2, then 3, to avoid systematically rejecting frequent pairs. We are aware that the modification of the formula is purely empirical. The low values of association ratio and cubic association ratio are assigned to pairs where the two lemmas appear seldom together and frequently separately as the pairs (système, terre) (*terrestrial system*, (code, signalisation) (*signalling code*), (bande, bruit) (*noise band*).

We are not going to present the pairs which own the high values of cubic association ratio because it is the same pairs that receive the highest values with the log-likehood coefficient. We simply remark that for the N_1 (PREP (DET)) N_2 structure, the values of the cubic association ratio take place between -1,01

and 21,34 for the STH corpus and between -4,51 and 25,20, and for the N ADJ structure, between -0,7 and 22,4 for the STH corpus and between -2,28 and 21,89 for the LBC corpus.

5.2.2 Log-Likehood coefficient

The log-likehood coefficient selects the same pairs as the cubic association ratio in the highest values. But, it is not defined if one of the lemma of the pair appears only in this pair. We can remark that when the log-likehood coefficient is not defined, the Yule coefficient (equation 5) takes its maximal value, namely 1, that the values of association ratio and Fager and MacGowan coefficient are among the highest and that one of the diversity associated with one of the lemma of the pair take the value 0. So, diversity is more precise thus it indicates if the two lemmas appears only together as (océan, indien) (*indian ocean*) (H_1=h_1=H_2=h_2=0), or, which of the two lemmas appears only with the other as (réseau, maillé) (*mesh network*) (H_2=h_2=0) where *maillé* appears only with *réseau* or for (codeur, idéal) (*ideal encoder*) (H_1=h_1=0) where the noun *codeur* appears only with the adjective *idéal*). Other examples are: (île, salomon) (*solomon islands*), (hélium, gazeux) (*helium gaz*), (suppresseur, écho) (*echo suppressor*), (retour, chariot) (*end-of-line*). These pairs for which the log-likehood coefficient is not defined collect numerous frozen compounds and collocations of the everyday French language.

N_1 (PREP (DET)) N_2

The values of the log-likehood coefficient take place between 0+ and 1 328 for the STH corpus and between 0+ and 5 738 for the LBC corpus. The range of values depends on the number of occurrences of the pair: the most frequent the pair is, the more important is the value of the log-likehood coefficient and this independently from the total number of extracted pairs. The pairs for which the log-likehood coefficient is not defined are 167 for the STH corpus and 169 for the LBC one. The following notations are used in the tables below:

Logl Log-Likehood coefficient

IM3 Cubic association ratio

Nbc Number of occurrences of the pair

Corpus	Pair of N_1 (PREP (DET)) N_2 structure	The most frequent pair sequence	**Logl**	IM3	Nbc
STH	(largeur, bande)	*largeur de bande* (197)	**1328**	21,34	223
	(température, bruit)	*température de bruit* (110)	**777**	20,13	126
	(bande, base)	*bande de base* (142)	**745**	19,88	145
	(amplificateur, puissance)	*amplificateur(s) de puissance* (137)	**728**	19,86	137
	(temps, propagation)	*temps de propagation* (93)	**612**	19,80	94
	(règlement, radiocommunication)	*règlement des radiocommunications* (60)	**521**	19,96	60
	(produit, intermodulation)	*produit(s) d'intermodulation* (61)	**458**	19,32	61
	(taux, erreur)	*taux d'erreur* (70)	**420**	18,61	70
	(mise, œuvre)	*mise en œuvre* (47)	**355**	18,60	47
	(télécommunication, satellite)	*télécommunication(s) par satellite* (88)	**353**	17,35	99
	(bilan, liaison)	*bilan(s) de liaison* (37)	**344**	17,99	55

Corpus	Pair of N_1 (PREP (DET)) N_2 structure	The most frequent pair sequence	**LogL**	IM3	Nbc
LBC	(canal, sémaphore)	*canal/canaux sémaphores* (1 188)	**5 738**	25,20	1 188
	(accusé, réception)	*accusé de réception* (558)	**3 983**	24,78	592
	(système, signalisation)	*système(s) de signalisation* (82)	**2 417**	23,05	85
	(complément, étude)	*complément d'étude* (242)	**1985**	23,74	245
	(point, sémaphore)	*point(s) sémaphore(s)* (677)	**1 822**	22,38	679
	(intervalle, temps)	*intervalle(s) de temps* (249)	**1 782**	23,19	251
	(trame, sémaphore)	*trame(s) sémaphore(s)* (354)	**1 444**	21,48	354
	(signal, fin)	*signal/signaux de fin* (385)	**1 407**	21,43	391
	(sou-système, utilisateur)	*sous-système utilisateur* (195)	**1 226**	22,10	195
	(bout, bout)	*bout en bout* (136)	**1 155**	22,58	137
	(contrôle, continuïté)	*contrôle(s) de continuïté* (171)	**1 116**	21,89	171

N ADJ

The values of Log-Likehood coefficient take place between 0+ and 2 934 for the STH corpus and between 0+ and 1 435 for the LBC one. The number of pairs for which this coefficient is not defined equals 147 for STH and 176 for LBC.

Corpus	Pair of N ADJ structure	**Logl**	IM3	Nbc	Corpus	Pair of N ADJ structure	**Logl**	IM3	Nbc
STH	(station, terrien)	**2934**	22,47	750	LBC	(équipement, terminal)	**1425**	21,63	275
	(débit, binaire)	**716**	19,45	134		(considération, général)	**1385**	21,38	256
	(accès, multiple)	**605**	19,10	105		(service, supplémentaire)	**1275**	21,15	340
	(voie, téléphonique)	**512**	18,52	118		(télégraphie, harmonique)	**1250**	21,89	152
	(liaison, montant)	**457**	17,50	88		(étude, ultérieur)	**1171**	21,52	169
	(liaison, descendant)	**408**	17,50	77		(caractère, graphique)	**1112**	21,09	196
	(secteur, spatial)	**341**	17,41	79		(entité, fonctionnel)	**999**	20,70	199
	(service, fixe)	**326**	17,42	66		(centre, international)	**964**	20,32	325
	(lobe, latéral)	**299**	17,93	40		(adresse, complet)	**874**	20,33	183
	(faisceau, hertzien)	**244**	17,22	35		(effet, local)	**865**	20,23	169
	(puissance, surfacique)	**205**	16,76	35		(station, mobile)	**855**	20,41	164

5.2.3 Fager and MacGowan coefficient

The high values of the Fager and MacGowan coefficient are mainly assigned to pairs whose two lemmas appear often together and seldom separately. This score is very close to association ratio, but accepts some frequent terms. The negative values of this score are assigned to pairs whose first lemma appears only with a second lemma which is used in numerous pairs, i.e. diversity of the first lemma is equal to zero, as for (prestataire, service) (*service provider*), (cheminement, information) (*information routing*), (ingénierie, trafic) (*traffic engineering*).
The values of Fager and MacGowan coefficient take place inside the maximal interval]-1,+1[.

N_1 (PREP (DET)) N_2

The values of the Fager and MacGowan coefficient take place between -0,310 and 0,849 for STH and -0,326 and 0,827 for LBC. The extracted pairs characterize terms of the telecommunication domain as *moteur d'apogée* (*apogee motor*), compounds of the current language as *accusé de réception* (*acknowledgement of*

receipt), frozen adverbs as *bout à bout* (*end-to-end*). We meet some frozen compounds which have already been identified by the association ratio as *aiguille d'une montre* (*clockwise*) and *glossaire du fascicule*. The following notations are used:

Fag Fager and MacGowan coefficient

Nbc Number of occurrences of the pair

Corpus	Pair of N_1 (PREP (DET)) N_2 structure	The most frequent pair sequence	**Fag**	Nbc
STH	(arséniure, gallium)	*arséniure de gallium* (11)	**0,849**	11
	(mémoire, tampon)	*mémoire(s) tampon(s)* (35)	**0,823**	35
	(égalité, droit)	*égalité de droits* (4)	**0,776**	5
	(règlement, radiocommunication)	*règlement des radiocommunications* (60)	**0,748**	60
	(batterie, accumulateur)	*batterie(s) d'accumulateurs* (11)	**0,711**	11
	(reportage, actualité)	*reportage(s) d'actualités* (2)	**0,711**	3
	(registre, décalage)	*registre à décalage* (7)	**0,698**	7
	(largeur, bande)	*largeur de bande* (197)	**0,691**	223
	(moteur, apogée)	*moteur d'apogée* (26)	**0,691**	28
	(aiguille, montre)	*aiguilles d'une montre* (2)	**0,646**	2
	(glossaire, fascicule)	*glossaire du fascicule* (2)	**0,646**	2

Corpus	Pair of N_1 (PREP (DET)) N_2 structure	The most frequent pair sequence	**Fag**	Nbc
LBC	(isolement, processeur)	*isolement de processeur* (24) *isolement du processeur* (11) *isolement des processeurs* (6)	**0,827**	41
	(complément, étude)	*complément d'étude* (242)	**0,731**	245
	(dos, dos)	*dos à dos* (3)	**0,711**	3
	(force, son)	*force des sons* (35)	**0,690**	38
	(accusé, réception)	*accusé de réception* (558)	**0,683**	592
	(compensation, dérive)	*compensation de dérive* (11)	**0,663**	13
	(microphone, charbon)	*microphone(s) à charbon* (35)	**0,658**	35
	(bout, bout)	*bout en bout* (136)	**0,640**	137
	(recommandation, série)	*recommandation de la série* (100)	**0,629**	123
	(tiers, octave)	*tiers d'octave* (14)	**0,625**	15
	(rapidité, modulation)	*rapidité(s) de modulation* (112)	**0,596**	117

N ADJ

The values of the Fager and MacGowan coefficient take place between -0,281 and 0,896 for STH and -0,306 and 0,885 for LBC. The retained pairs are either terms of the telecommunication domain as *lobes latéraux* (*side lobes*) or *station terrienne* (*earth station*) , or proper nouns as *île Salomon* (*salomon islands*) or *océan indien* (*indian ocean*) , or composed subordinated conjunction as *compte tenu* (*due to*), or collocations as *période probatoire*. Only the candidate *étude ultérieure* do not convince us, but it could eventually be considered as a collocation.

Corpus	Pair of N ADJ structure	**Fag**	Nbc	Corpus	Pair of N ADJ structure	**Fag**	Nbc
	(compte, tenu)	**0,896**	23		(télégraphie, harmonique)	**0,885**	152
	(station, terrien)	**0,842**	750		(compte, tenu)	**0,852**	66
	(publication, anticipée)	**0,823**	8		(polynôme, générateur)	**0,849**	11
	(stabilisation, triaxiale)	**0,811**	7		(faute, matériel)	**0,828**	38
	(rafraîchissement, conditionnel)	**0,811**	7		(période, probatoire)	**0,800**	75
STH	(île, salomom)	**0,750**	4	LBC	(assemblée, plénière)	**0,796**	6
	(distorsion, intersymbole)	**0,737**	6		(accord, bilatéral)	**0,787**	80
	(couple, perturbateur)	**0,715**	7		(signe, diacritique)	**0,755**	31
	(hélium, gazeux)	**0,711**	3		(océan, indien)	**0,750**	4
	(lobe, latéral)	**0,699**	40		(étude, ultérieur)	**0,735**	169
	(atmosphère, clair)	**0,693**	12		(prise, simultané)	**0,683**	78

To conclude, after the examinationof the sorts proposed by these three scores, we wish to only retain the Log-Likehood coefficient. Indeed, the Fager and MacGowan one gives to much importance to pairs where the two lemmas appear often together and seldom separately; the cubic association ratio gives good results but, as its formula relies on an empiristic study, we repel to retain it. The log-likehood coefficient, proposed by [Dunning, 1993], owns the following properties:

- it is a real statistical test,
- it proposes a sorting which takes into account frequency of the pairs,
- it reacts as well on a small corpus as on a big one,
- it is not defined for a number of pairs which do not accept diversity on one of them items: these pairs which are systematically retain by other association criteria can own a frozen character and belong more often the the general language than to technical one. It is so important that they could be isolated.

We give in annex B a sample of the sorting provided by the Log-Likehood coefficient only for the STH corpus but for our two syntactic structures. We include the list of pairs for which this coefficient is not defined; it is easy to state that half of them are terms. These lists include the pairs, as well as their regular expressions when the pair has been encountered only in one or two forms. For the other pairs, no regular expression is indicated. In order to give an element of comparison, we indicate the frequency of the pair, the value of the association ratio at square and the value of the Fager and MacGowan coefficient. We have also indicated the values of the normalized diversity for reason that we are going to explain.

5.3 Diversity

The normalized diversity provides interesting informations about the distribution of the pair lemmas in the set of pairs. A lemma with a high diversity means that it appears in several pairs in equal proportion; conversely, a lemma which

appear only in one pair owns a zero diversity (minimal value) and this, whatever is the frequency of the pair. We have already seen a few examples of pairs characterized by a zero diversity on one of their lemma: these pairs receive high values of association ratio, maximal value of Yule coefficient and a non-definite value of Log-Likehood coefficient. We give two new examples: considering the adjective *anormal*, in the N ADJ syntactic pattern, only the noun *fonctionnement* (*malfunction*) has been encountered and we have $H_2 = H_{anormal} = 0$; the noun *fonctionnement* can appear with other adjectives. Conversely, considering the noun *éclaircissement*, in the N ADJ syntactic pattern, only the adjective *nécessaire* (*necessary clarification*) has been encountered and we have $H_1 = H_{éclaicissement} = 0$; of course, this does not imply that the adjective *nécessaire* does not appear with other nouns. Diversity has been computed for each lemma which appear in a pair in a given position.

$\mathbf{N_1}$ (PREP (DET)) $\mathbf{N_2}$

1. Diversity applied to N_1

   - For the STH corpus, the tenth high values are assigned to the following nouns: *fonctionnement* , *fonction*, *moyen*, *type*, *partie*, *compte*, *cas*, *raison* , *exemple*, *signal*, *caractéristique*.
   - For the LBC corpus, the tenth high values are assigned to the following nouns: *fonctionnement*, *cas*, *partie*, *signalisation*, *utilisation*, *moyenne*, *compte*, *moyen*, *exemple*, *procédure*.

   Although the high values of the preceding scores do not characterize the same pairs through the two corpora, it is interesting to remark that highest values of diversity are shared by the same nouns through the two corpora of different size. Further more, high values of diversity applied to N_1 characterize either nouns that appear inside a composed preposition as *en fonction de*, *au moyen de*, *en raison de*, *en cas de*, *en tenant compte de* or *compte tenu de*, or quantifiers as *nombre de*, or classifiers as *partie de*, *type de*.

2. Diversity applied to N_2

   - For the STH corpus, the tenth high values are assigned to the following nouns: *système*, *station*, *réseau* (*network*), *signal*, *service*, *équipement*, *fonctionnement* (*operation*), *antenne* (*antenna*), *fréquence*, *niveau*.
   - For the LBC corpus, the tenth high values are assigned to the following nouns: *système*, *service*, *fonctionnement* (*operation*), *équipement*, *niveau*, *fonction*, *réseau* (*network*), *commutateur*, *circuit*, *signaleur*.

   Here again, although the corpus do not have the same size, the list are nearly identical. The nouns with a high diversity applied on N_2 characterize the technical domain (except some of them as *fonction* or *niveau*); we could call them the key-words of the telecommunication domain.

N ADJ

1. Diversity applied to ADJ
   Diversity applied to the adjective for the N ADJ structure allow us to identify adjectives which do not take part to base-terms as *nécessaire* (*necessary*), *suivant* (*following*), *important*, *différent* (*various*), *tel* (*such*), etc. Here again, it is interesting to remark that they are the same adjective which appear in the two lists for the two corpora.

   - For the STH corpus, the tenth high values are assigned to the following adjectives: *nécessaire* (*necessary*), *suivant* (*following*), *différentiel*, *numérique*, *général*, *important*, *supplémentaire*, *particulier*, *relatif*, *différent* (*various*).
   - For the LBC corpus, the tenth high values are assigned to the following adjectives: *suivant* (*following*), *différentiel*, *différent* (*various*), *correspondant*, *particulier*, *possible*, *spécifique*, *tel* (*such*), *nécessaire* (*necessary*), *normal*.

2. Diversity applied to N

   High values of diversity applied to the noun for the N ADJ structure could be seen as key-words of the domain:

   - For the STH corpus, the tenth high values are assigned to the following nouns: *valeur*, *fonctionnement*, *système*, *réseau* (*network*), *caractéristique*, *équipement*, *niveau*, *signal*, *antenne* (*antenna*), *satellite*.
   - For the LBC corpus, the tenth high values are assigned to the following nouns: *signal*, *fonctionnement*, *façon*, *signalisation*, *fonction*, *système*, *valeur*, *mesure*, *caractéristique*, *méthode*.

In [Daille, 1993], we have shown how it is possible to use the results provided by the diversity to filter bad candidates in the sorting proposed by the frequency. With the sorting proposed by the log-likehood coefficient, the pairs which share a high diversity do not receive high values. But, the informations provide by the diversity could be useful for middle values of log-likehood coefficient; it is why we decided to give here these high values of diversity. In future work, we will investigate how to incorporate the nice results provided by diversity into an automatic extraction algorithm.

5.4 Distance Measures

The distance measures bring interesting informations which concern the morphosyntactic variations of the base-terms, but they don't allow to take a decision upon the status of term or non-term of a candidate. A pair which has no distance variation, whatever is the distance, is or is not a term; we give now some examples of pairs which have no distance variations and which are not terms: *paire de signal* (*a pair of signal*), *type d'antenne* (*a type of antenna*), *organigramme de la figure* (*diagram of the figure*), etc.

$\mathbf{N_1}$ (PREP (DET)) $\mathbf{N_2}$

We illustrate below how the distance measures allow to attribute to a pair its elementary type automatically, for example, either N_1 N_2, N_1 PREP N_2, N_1 PREP DET N_2, or N_1 ADJ PREP (DET) N_2 for the general N_1 (PREP (DET)) N_2 structure.
In the examples below, we use the following notations: *Dist*, the mean of the number of items which occur between the two lemmas, *Var*, the variance of this number of items and *MDist*, the mean of the number of main items which occur between the two lemmas. We indicate for each pairs the flexions and the variations that have been observed.

1. **Pairs with no distance variation** $V(X) = 0$
   70 % of the set of pairs for the STH corpus and 65 % for the LBC one do not accept distance variations.

   (a) $\mathbf{N_1}$ $\mathbf{N_2}$: $Dist = 2$ $MDist = 2$
   - *liaison sémaphore*, *liaisons sémaphores* (*common signalling link(s)*)
   - *canal support*, *canaux support*, *canaux supports* (*bearer channel*)

   (b) $\mathbf{N_1}$ PREP $\mathbf{N_2}$: $Dist = 3$ $MDist = 2$
   - *accusé(s) de réception* (*acknowledgement of receipt*)
   - *refroidissement à air*, *refroidissement par air* (*cooling by air*)

   (c) $\mathbf{N_1}$ PREP DET $\mathbf{N_2}$: $Dist = 4$ $MDist = 2$
   - *sensibilité au bruit* (*susceptibility to noise*)
   - *reconnaissance des signaux* (*signal recognition*)

   (d) $\mathbf{N_1}$ ADJ PREP $\mathbf{N_2}$: $Dist = 4$ $MDist = 3$
   - *réseau local de lignes*, *réseaux locaux de lignes* (*local line network(s)*)
   - *service fixe par satellite* (*fixed-satellite service*)

2. **Pairs with distance variations** $V(X) \neq 0$
   30 % of the set of pairs for the STH corpus and 35 % for the LBC one accept distance variations. These scores confirm our linguistic study of the terms of the telecommunication domain and those of [Jacquemin, 1991]: terms of a technical domain are not frozen compounds:

   - (demande, trafic)
     demande de trafic
     demandes en trafic
     demande réelle en trafic
   - (liaison, satellite)
     liaison par satellite, *liaisons par satellite*

liaisons (très rapides + numériques + téléphoniques nationales) par satellite
liaisons numériques par satellites
liaisons satellite
liaisons entre satellites

- (ligne, abonné)
  ligne d'abonné, *lignes d'abonné*
  ligne de l'abonné, *lignes de l'abonné*
  ligne d'abonnés, *lignes des abonnés*
  ligne(s) (téléphonique(s) + numériques(s) + analogique(s)) d'abonné
  ligne(s) (numérique(s) + analogique(s)) de l'abonné
  lignes et services d'abonné
- (signal, fin)
  signal de fin , *signaux de fin*
  signal (local + national + valide + périodique) de fin
  signal émis à des fins
  signal numérique utilisé à des fins
  The two last occurrences above illustrate the problem described at the beginning of this report: we have no assurance that a pair collects co-occurrences that refer to an unique concept, or as this example, co-occurrences which are all valid ones.

The modifiers that could be inserted inside a base-term are not numerous. Recording these modifiers, as the other possible alterations of the structure of the base-term and including the various flexions encountered, is easily bring about automatically. These lexical informations are present under a pair and could be directly integrated in a dictionary.

N ADJ

The pairs of N ADJ structure that accept internal modification are less numerous than the N_1 (PREP (DET)) N_2. This comes from the fact that we did not allow inserted adjectives in this structure. 13 % of the set of pairs for the STH corpus and 11 % for the LBC one accept distance variations.

1. **Pairs with no distance variation** $V(X) = 0$
   These pairs are for 99 % of structure N ADJ as *satellite artificiel*, *rayonnement solaire*, *cadre supérieur*. The other possible structures of fixed length are:
   (a) N *non* ADJ
   - *amplificateur(s) non linéaire(s)*
   - *numéro non valable*
   - *rayonnement(s) non essentiel(s)*

   (b) N *adv* ADJ
   - *filtre passe-bas idéal*

- *onde avant absente, onde arrière absente*

(c) N ADJ CONJ ADJ

- *idéogrammes chinois et japonais*
- *ondes métriques et décimètriques*
- *organisme(s) scientifique(s) ou industriel(s)*

A few pairs as (codage, décimal) have only occurrences observed in an attributive structure as *codage est décimal.*

2. **Pairs with distance variations** $V(X) \neq 0$
   The pairs where the adjective appears either as an epithet or as an attribute represent 20 % of the set of pairs for the STH corpus and 30 % for the LBC one. Thus, the distance variation the most frequent is caused by the insertion of an adverb.

   - (circuit, numérique)
     circuit numérique, circuits numériques
     circuits entièrement numériques
     circuits analogiques et numériques
     circuits sont numériques
   - (effet, local)
     effet local
     effet purement local
     effet uniquement local
   - (ligne, téléphonique)
     ligne téléphonique, lignes téléphoniques
     ligne (téléphonique), lignes (téléphoniques)

In the list of pairs which are given in Annex B, we do not indicate the values of the distance measures. But we precise for several pairs, the regular expression corresponding to the different sequences which have been encountered.

6 Conclusion

To conclude with the examination of the statistical scores that have been selected by graphical evaluation, we affirm that the frequency of a pair is a good indicator of its terminological character. The problem with the frequency is that it does not allow to isolate rare terms and that noise comes very quickly, even if this noise could be reduced using results of normalized diversity. Between frequency and the three association criteria, we have chosen to only retain the log-likehood coefficient for its following properties: its nature (a well grounded statistical test), its general tendency to make a good use of the frequency of the pairs, its good behavior whatever the corpus size and the indefinite value that it assigns to pairs with a null diversity. Nevertheless, even if the log-likehood coefficient behaves itself as the best statistical score, the sorting that it provides includes some noise that we now analyze:

1. some pairs have been produced because of a wrong tagging:
   - (pas, traduction) (*pas de traduction*) where the negative adverb *pas* is tagged as a noun.
2. a few pairs whose one of the lemma is a term of lengh 1 built with an hyphen appear in double in the list; it is the fault of the lemmatizer which do not correctly lemmatize compounds with an hyphen:
   - (sou- système, utilisateur), (sou-systèmes, utilisateur)
3. some of the pairs are not of nominal type but are:
   - adverbs or sub-sequences of a prepositional phrase of adverbial type:
     - (bout, bout) (*bout en bout*)
     - (titre, exemple) (*à titre d'exemple*),
     - (plupart, cas) (*dans la plupart des cas*)
     - (heure, actuelle) (*à l'heure actuelle*)
   - measure units:
     - (ko, bit) (*ko bits*)
4. a most important number of pairs refer to a sub-sequence of a term of length ≥ 3 . This noise straightly results from the problem of overcomposition and modification presented in [Daille, 1994].
   - (rapport, porteur) for example sub-sequence of the term *rapport porteuse/bruit*,
   - (type, limiteur) sub-sequence of*linéariseur de type à limiteur*,
   - (augmentation, espacement) sub-sequence of *augmentation de l'espacement angulaire*
   - (accès, étalement) sub-sequence of *accès multiple par étalement du spectre*
   - (service, fixe) sub-sequence of *service fixe par satellite*,
   - (circuit, fictif) sub-sequence of *circuit fictif de référence*,
   - (bande latérale) sub-sequence of *bande latérale unique*.

Conversely, having accepted the insertion of modifiers inside the base-term structures implies to identify a few terms of length 3. For example, the pair (service, satellite) represents the term *service fixe par satellite* (*fixed satellite service*) and the pair (accès, répartition) the term *accès multiple par répartition*. We should also precise that a few base-terms do not belong to the specific domain of the telecommunication but rather to the current language as (feuille, papier) (*feuille(s) de papier*), or (mise, page) (*mise en page*). These results are not perfect, but noise is considerably minimized compare to the lists obtained only with statistical methods, i.e. without linguistic filters. The choice to only retain the log-likehood coefficient implies the no selection of pairs which are base-terms. However, whatever the retained score is, we would had faced

the same problem, i.e. some base-terms would have been forgiven, without obtaining a sorting of high accuracy.
The objective of this work was to extract the base-term and its various variants. We have obtained the following ones :

- **Orthographic variants**:
  Under a pair, we find: the flexions and the graphics of the base-term encountered in the corpus. For the N_1 N_2 structure with allows an optional hyphen, we have considered that the base-terms which accept the two forms are variants of the term of length 1, i.e. built with an hyphen. We have checked which terms accept the two forms : there are only two terms in the STH corpus: *écart-type/ écart type* and *schéma-type/schéma type*, and only three in the LBC one: *écart-type/ écart type*, *mémoire-tampon/mémoire tampon* and *mode-paquet/mode paquet*. This result is very interesting: the hyphen in the telecommunication domain owns a frozen character which should not be the same in another technical domain.

- **Morphosyntactic variants** :
  We find under a pair of N_1 (PREP (DET)) N_2 structure, the variants that imply a simplication or a complication of the structure and the switches of preposition. On the contrary, we did not identify the synonymic relations; those should be identified on the bilingual terminology extraction.

- **Elliptical variants**:
  We did not foresee the extraction of elliptical variants. However, having extracted a few terms of length 3 of N_1 ADJ PREP N_2 structure has allowed to list some variants as, for example, the elliptical variant *service par satellite* of the term *service fixe par satellite*.

- **abbreviations**:
  The main abbreviations of our corpora have been manually extracted. They should be added to the other variants.

References

[Bourigault, 1994] Bourigault (Didier). – *Acquisition de terminologie.* – PhD thesis, EHESS, 1994.

[Brown *et al.*, 1988] Brown (Peter F.), Cocke (John), Pietra (Stephen A. Della), Pietra (Vincent J. Della), Jelinek (Frederik), Mercer (Robert L.) et Roossin (Paul S.). – A statistical approach to language translation. *In: Proceeding of the 12^{th} International Conference on Computational Linguistics (Coling-88).* – Budapest, Hungary, Août 1988.

[Calzolari and Bindi, 1990] Nicoletta Calzolari and Remo Bindi. 1990. Acquisition of lexical information from a large textual Italian corpus. In *Proceedings of the Thirteenth International Conference on Computational Linguistics*, Helsinki, Finland.

[Church et Hanks, 1989] Church (Kenneth Ward) et Hanks (Patrick). – Word association norms, mutual information, and lexicography. *In : Proceeding of the 27th Annual Meeting of the ACL.* – Vancouver, Canada, Juin 1989.

[Church et Hanks, 1989] Church (Kenneth Ward) et Hanks (Patrick). – Word association norms, mutual information, and lexicography. *In : Proceeding of the 27th Annual Meeting of the ACL.* – Vancouver, Canada, Juin 1989.

[Church and Hanks, 1990] Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, vol. 16, n^o 1, pp. 22–29.

[Daille, 1993] Daille (Béatrice). – Extraction automatique de terminologie monolingue. *In : Colloque Informatique et Langue Naturelle.* Nantes, December 93.

[Daille, 1994] Béatrice Daille. 1994. *Approche mixte pour l'extraction automatique de terminologie : statistique lexicale et filtres linguistiques.* PhD thesis, University Paris 7, France.

[Daille *et al.*, 1994] 1994. Béatrice Daille, Éric Gaussier and Jean-Marc Langé. Towards Automatic Extraction of Monolingual and Bilingual Terminology. *COLING-94*, Kyoto, Japon.

[Dunning, 1993] Dunning (Ted). – Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, vol. 19, n^o 1, March 1993.

[Enguehard, 1992] Enguehard (Chantal). – *ANA, Apprentissage Naturel Automatique d'un réseau sémantique.* – PhD thesis, University of technology Compiègne, 1992.

[Gale et Church, 1991b] Gale (William A.) et Church (Kenneth W.). – Concordances for parallel texts. *In : Proceedings of the Seventh Annual Conference of the UW Centre for the New OED and Text Research, Using Corpora*, pp. 40–62. – Oxford, U.K., 1991.

[Gaussier, 1994] Gaussier (Éric). – *Introduction des probabilités dans le module de transfert en traduction automatique.* – PhD thesis, University Paris VII, 1994.

[Jacquemin, 1991] Christian Jacquemin. 1991. *Transformations des noms composés.* PhD thesis, University Paris 7, France.

[Lafon, 1984] Pierre Lafon. 1984. *Dépouillements et Statistiques en Lexicométrie*, Genève, Slatkine, Champion.

[Gaussier *et al.*, 1992] Gaussier (Éric), Langé (Jean Marc) et Meunier (Frederic). – Towards bilingual terminology. *In : Proceedings of ALLC/ACH Conference.* – Oxford, England, 1992.

[Shannon, 1948] C. Shannon. 1948. The mathematical theory of communication. *Bell Systems Technical Journal*, 27.

[Smadja and McKeown, 1990] Frank A.Smadja and Kathleen R.McKeown. 1990. Automatically extracting and representing collocations for language generation. In:*Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, pp. 252–259. – Pittsburgh.

A French Tags

AAAA Punctuation

ADJEFP adjective feminine plural

ADJEFS adjective feminine singular

ADJEMP adjective masculine plural

ADJEMS adjective masculine singular

ADJIFP adjective indefinite feminine plural

ADJIFS adjective indefinite feminine singular

ADJIMP adjective indefinite masculine plural

ADJIMS adjective indefinite masculine singular

ADVE adverb

AUXA auxiliary *avoir*

AUXA1 auxiliary *avoir* person 1. singular

AUXA2 auxiliary *avoir* person 2. singular

AUXA3 auxiliary *avoir* person 3. singular

AUXA4 auxiliary *avoir* person 1. plural

AUXA5 auxiliary *avoir* person 2. plural

AUXA6 auxiliary *avoir* person 3. plural

AUXE auxiliary *être*

AUXE1 auxiliary *être* person 1. singular

AUXE2 auxiliary *être* person 2. singular

AUXE3 auxiliary *être* person 3. singular

AUXE4 auxiliary *être* person 1. plural

AUXE5 auxiliary *être* person 2. plural

AUXE6 auxiliary *être* person 3. plural

CCOO coordinating conjunction

CHIF number

CSUB subordinating conjunction

DETRFP feminine plural determiner

DETRFS feminine singular determiner

DETRMP masculine plural determiner

DETRMS masculine singular determiner

DINTFP feminine plural wh-determiner

DINTFS feminine singular wh-determiner

DINTMP masculine plural wh-determiner

DINTMS masculine singular wh-determiner

NE negation

NPRO proper noun

PAS *pas*, *plus*

PAU contracted preposition *au* singular

PAUX contracted preposition *aux* plural

PDEA preposition *de*, *d'*, *à* and *es*

PDES contracted preposition *des* plural

PDETFP feminine plural demonstrative pronoun

PDETFS feminine singular demonstrative pronoun

PDETMP masculine plural demonstrative pronoun

PDETMS masculine singular demonstrative pronoun

PINDFS feminine singular indefinite pronoun

PINDFP feminine plural indefinite pronoun

PINDMS masculine singular indefinite pronoun

PINDMP masculine plural indefinite pronoun

PINTFS feminine singular wh-pronoun (question)

PINTFP feminine plural wh-pronoun (question)

PINTMS masculine singular wh-pronoun (question)

PINTMP masculine plural wh-pronoun (question)

PPASFP feminine plural past-participle

PPASFS feminine singular past-participle

PPASMP masculine plural past-participle

PPASMS masculine singular past-participle

PPER1 person 1. singular personal pronoun

PPER2 person 2. singular personal pronoun

PPER3F person 3. feminine singular personal pronoun

PPER3M person 3. masculine singular personal pronoun

PPER4 person 1. plural personal pronoun

PPER5 person 2. plural personal pronoun

PPER6F person 3. feminine plural personal pronoun

PPER6M person 3. masculine plural personal pronoun

PPOBFP person 1. feminine plural objective personal pronoun

PPOBFS person 1.feminine singular objective personal pronoun

PPOBMP person 1. masculine plural objective personal pronoun

PPOBMS person 1.masculine singular objective personal pronoun

PPRE present participle

PREFFP feminine plural reflexive pronoun

PREFFS feminine singular reflexive pronoun

PREFMP masculine plural reflexive pronoun

PREFMS masculine singular reflexive pronoun

PRELFP feminine plural wh-pronoun (relative)

PRELFS feminine singular wh-pronoun (relative)

PRELMP masculine plural wh-pronoun (relative)

PRELMS masculine singular wh-pronoun (relative)

PREP preposition

PREPMS contracted preposition *du* masculine singular

SUBSFP feminine plural common noun

SUBSFS feminine singular common noun

SUBSMP masculine plural common noun

SUBSMS masculine singular common noun

VERB1 verb person 1 singular

VERB2 verb person 2 singular

VERB3 verb person 3 singular

VERB4 verb person 1 plural

VERB5 verb person 2 plural

VERB6 verb person 3 plural

VINF infinitive verb

XFAMIL family name

XPAYFP feminine plural country noun

XPAYFS feminine singular country noun

XPAYMP masculine plural country noun

XPAYMS masculine singular country noun

XPREF feminine first-name

XPREM masculine first-name

XSOC enterprise name

XVILLE City name

YAAA punctuation

ZTRM end-of-line

B Sorting of the pairs obtained with the log-likehihood coefficient

STH Corpus

N_1 (PREP (DET)) N_2

Sorting of Log-likehood coefficient

Index	N_1	N_2	R. expression	NC	LOG	FAG	IM3	h1
514	largeur	bande		223	1327	0.691	21.34	0.38
5110	température	bruit		126	777	0.592	20.13	0.65
6	bande	base	*(bande + bandes) de base*	145	745	0.515	19.88	2.19
7194	amplificateur	puissance	*(amplificateur + amplificateurs) de puissance*	137	727	0.517	19.86	1.37
2937	temps	propagation		94	611	0.611	19.80	1.80
3904	règlement	radiocommunication	*règlement des radiocommunications*	60	521	0.803	19.96	0.49
2604	produit	intermodulation	*(produit + produits) d' intermodulation*	61	457	0.626	19.31	0.63
3256	taux	erreur	*taux d' erreur*	70	420	0.448	18.61	1.02
1	mise	oeuvre	*mise en oeuvre*	47	355	0.563	18.60	2.16
2	télécommunication	satellite		99	353	0.228	17.35	1.05
678	bilan	liaison		55	349	0.390	17.99	0.26
1297	mémoire	tampon	*mémoires tampons + mémoire tampon*	35	343	0.823	19.30	0.70
3810	concentration	conversation		39	315	0.630	18.71	1.08
761	diagramme	rayonnement	*(diagramme + diagrammes) de rayonnement*	40	306	0.578	18.51	1.47
7383	angle	site		37	300	0.622	18.59	1.69
7776	propagation	groupe	*propagation de groupe*	38	292	0.574	18.43	1.52
3423	correction	erreur		46	280	0.352	17.49	0.75
8617	moteur	apogée		28	253	0.683	18.49	1.22
1200	titre	exemple	*titre d' exemple*	29	252	0.645	18.40	1.08
5672	dispersion	énergie	*dispersion d' énergie*	33	249	0.453	17.72	0.38
6116	répéteur	satellite		76	246	0.177	16.41	1.44
3324	réduction	puissance		52	245	0.261	16.84	1.41
5718	assignation	demande	*assignation à la demande*	30	233	0.534	17.91	1.31
9774	capacité	trafic		53	230	0.262	16.75	2.90
3	bande	fréquence		89	223	0.187	16.50	2.19
3649	puissance	sortie		57	222	0.241	16.54	2.99
8616	transmission	donnée		65	221	0.216	16.46	3.06
9349	objectif	qualité		34	216	0.352	17.08	1.06

Index	N_1	N_2	R. Expression	NC	LOG	FAG	IM3	h1
8411	diamètre	antenne		39	194	0.200	16.00	0.71
5746	alimentation	énergie	*alimentation (en + d') énergie*	33	186	0.304	16.57	2.06
2546	code	convolution	*(code + codes) à convolution*	24	183	0.416	16.84	2.71
8854	schéma	principe	*(schéma + schémas) de principe*	23	180	0.471	17.18	2.59
8972	unité	voie		42	180	0.219	16.05	2.52
9472	accusé	réception	*(accusé + accusés) de réception*	28	179	0.240	16.10	0.15
8135	suppresseurs	écho	*suppresseurs d' écho*	25	175	0.375	16.85	0.64
4268	multiplication	circuit		25	171	0.307	16.47	0.54
2852	antenne	station		62	171	0.159	15.70	2.74
6119	multiplexage	répartition	*multiplexage par répartition*	23	170	0.412	16.96	1.54
9381	accès	répartition	*accès multiple (par + à) répartition*	28	170	0.327	16.48	2.66
824	canal	sémaphore	*canal sémaphore + canaux sémaphores*	26	166	0.344	16.57	2.27
7011	couplage	amplification		21	165	0.453	17.09	1.32
6584	programme	télévision	*(programme + programmes) de télévision*	28	163	0.267	16.17	1.77
7111	guide	onde		20	153	0.345	16.50	0.30
5881	réseau	satellite		97	153	0.133	15.78	2.94
9163	système	signalisation		85	150	0.143	15.63	3.66
466	gain	antenne		46	149	0.153	15.32	2.54
5218	zone	couverture	*(zone + zones) de couverture*	22	147	0.348	16.37	2.04
5708	maintien	position	*maintien en position*	18	145	0.449	16.86	1.35
6284	service	radiodiffusion		28	144	0.236	15.47	3.23
1269	traitement	signal		39	143	0.159	15.25	2.25
1365	liaison	satellite		82	141	0.124	15.43	2.92
2925	modulation	delta	*modulation delta*	19	137	0.325	15.84	2.52
3935	méthode	modulation		29	134	0.204	15.37	3.04
2241	train	bit	*(train + trains) de bits*	21	132	0.282	15.90	1.37
8913	facteur	qualité	*(facteur + facteurs) de qualité*	27	131	0.206	15.37	2.50
2777	rapport	porteur		42	129	0.150	14.98	**3.38**
7915	durée	vie	*durée de vie*	16	127	0.411	16.34	2.85
5292	station	référence		36	126	0.165	14.91	3.51
6657	amplification	puissance	*amplification de puissance*	23	126	0.150	14.95	0.57
9097	système	alimentation	*(système + systèmes) d' alimentation*	47	123	0.142	14.76	3.66
8485	puissance	entrée		38	123	0.153	14.84	2.99
4246	pourcentage	temps		20	122	0.232	15.51	1.49
7480	batterie	accumulateur	*(batterie + batteries) d' accumulateurs*	11	121	0.715	17.46	0.75
148	code	bloc		21	120	0.254	15.41	2.71
1028	mise	place	*mise en place*	18	119	0.291	15.50	2.16
7377	tube	onde	*(tube + tubes) à ondes*	19	118	0.249	15.53	1.31
4771	trame	sémaphore	*trame sémaphore + trames sémaphores*	17	116	0.292	15.78	1.73
954	module	interface		16	115	0.303	15.85	1.33
554	étalement	spectre	*étalement (de + du) spectre*	13	114	0.379	16.24	0.26
7861	centre	commutation	*(centre + centres) de commutation*	27	113	0.171	14.79	2.79
7813	liaison	connexion	*(liaison + liaisons) de connexion*	24	112	0.192	14.72	2.92
248	faisceau	antenne		32	112	0.120	14.49	2.40
6425	modulation	déplacement	*modulation par déplacement*	15	107	0.282	15.14	2.52
7329	radiodiffusion	satellite		26	105	0.057	13.73	0.31
1431	déplacement	phase		16	104	0.217	15.19	0.97
337	orbite	transfert		15	104	0.303	15.59	1.22
3601	voie	retour	*(voie + voies) de retour*	18	102	0.219	14.69	3.68
4154	distance	coordination	*distance de coordination*	15	102	0.271	15.45	1.77
1699	commande	orientation	*commande d' orientation*	16	102	0.262	15.10	3.35
4326	bruit	intermodulation	*bruit d' intermodulation*	28	101	0.147	14.36	3.39
100	bout	bout	*bout en bout*	11	101	0.475	16.46	1.12
239	intervalle	temps	*(intervalle + intervalles) de temps*	17	99	0.187	14.85	1.51
453	rayonnement	antenne		22	97	0.098	14.08	1.30
7592	démodulateur	seuil	*(démodulateur + démodulateurs) à seuil*	13	97	0.307	15.60	1.65
7985	bit	contrôle	*(bit + bits) de contrôle*	18	97	0.207	14.76	3.00
2128	point	vue	*point de vue*	18	96	0.207	14.64	3.36
8438	système	satellite		108	95	0.110	15.32	3.66

Index	N_1	N_2	R. expression	NC	LOG	FAG	IM3	h1
5971	réseau	terre		44	95	0.111	14.21	2.94
7330	service	satellite		66	95	0.095	14.54	3.23
6696	équipement	multiplexage		24	92	0.148	14.04	3.61
9671	entrée	récepteur	*entrée du récepteur*	16	91	0.212	14.67	2.93
4781	poursuite	échelon	*poursuite par échelons*	11	90	0.376	15.73	2.00
1271	interface	terre	*(interface + interfaces) de terre*	23	90	0.112	14.00	2.40
8936	flux	donnée		19	89	0.113	14.07	1.43
4697	répartition	fréquence	*répartition en (fréquence + fréquences)*	23	88	0.084	13.77	1.48
3602	erreur	pointage		18	88	0.173	14.31	3.24
2203	signalisation	enregistreur	*signalisation (entre + d') enregistreurs*	15	87	0.192	14.04	3.80
6198	train	impulsion	*train d' impulsions*	13	86	0.254	14.98	1.37
6402	équipement	télécommunication	*(équipement + équipements) de télécommunication*	38	86	0.106	13.92	3.61
8344	centre	contrôle	*centre de contrôle*	18	86	0.169	14.18	2.79
5705	satellite	télécommunication		30	85	0.107	13.82	3.77
4979	excursion	fréquence		20	83	0.073	13.58	1.04
4284	registre	décalage	*registre à décalage*	7	83	0.698	16.89	0.38
3040	affaiblissement	espace	*affaiblissement en espace*	13	82	0.232	14.67	2.45
2879	boucle	verrouillage	*boucle à verrouillage*	9	80	0.404	15.64	2.04
7112	onde	polarisation		14	79	0.166	14.30	2.28
3467	récupération	porteur		16	79	0.101	13.79	1.29
4166	modulation	fréquence	*modulation de fréquence*	33	79	0.085	13.60	2.52
4197	rapport	signal		35	78	0.093	13.65	**3.38**
8474	système	couplage	*(système + systèmes) de couplage*	20	76	0.124	13.19	3.66
9636	chaîne	amplification		14	76	0.179	14.13	2.19
7074	longueur	onde	*(longueur + longueurs) d' onde*	14	76	0.156	14.10	2.29
6223	exploitation	réseau	*exploitation (du réseau + des réseaux)*	21	74	0.086	13.42	2.29
6681	algorithme	décodage	*(algorithme + algorithmes) de décodage*	10	74	0.260	14.92	1.82
8487	signal	entrée		30	73	0.100	13.42	3.95
9109	modèle	référence	*modèle de référence*	12	72	0.136	14.03	1.37
8307	densité	puissance		17	72	0.078	13.35	1.30
6112	système	télécommunication		46	71	0.093	13.68	3.66
975	conduit	référence		11	71	0.133	14.07	0.86
9076	débit	information		15	71	0.132	13.68	2.61
4987	matrice	commutation	*(matrice + matrices) de commutation*	12	70	0.129	13.91	0.98
5415	tube	hyperfréquence		11	70	0.204	14.32	1.31
7340	égaliseur	temps		11	69	0.124	13.92	1.04
1004	table	erlang	*tables d' erlang*	6	69	0.613	16.42	0.41
3072	terre	espace	*terre vers espace*	11	69	0.204	14.21	2.40
6176	stabilisation	rotation	*stabilisation par rotation*	8	68	0.270	15.02	0.64
8768	réduction	débit	*réduction (de + du) débit*	15	68	0.131	13.51	1.41
223	axe	faisceau		13	68	0.132	13.71	2.40
1645	point	multipoint	*point à multipoint*	11	67	0.195	13.79	3.36
1413	effet	champ	*effet de champ*	10	67	0.228	14.20	3.58
6243	service	exploration	*service d' exploration*	12	66	0.161	13.29	3.23
4789	poursuite	mono-impulsion	*poursuite mono-impulsion*	8	66	0.325	14.93	2.00
995	verrouillage	phase	*verrouillage de phase*	10	65	0.124	13.87	0.90
152	mot	code	*(mot + mots) de code*	10	65	0.151	14.00	1.43
7790	générateur	secours	*(génératrice + génératrices) de secours*	10	64	0.182	14.10	2.12
5779	réseau	communication		29	64	0.088	13.09	2.94
3761	niveau	brouillage		20	64	0.100	13.04	3.60
4664	modulation	amplitude	*modulation d' amplitude*	14	62	0.130	13.22	2.52
4914	synchronisation	trame		12	62	0.132	13.51	2.59
4251	contour	coordination	*(contour + contours) de coordination*	8	62	0.177	14.32	0.64
1361	liaison	terre		31	62	0.080	13.08	2.92
4623	ligne	alimentation		14	62	0.103	13.20	2.69
9683	chaîne	émission		18	61	0.080	12.93	2.19
1345	affaiblissement	pluie		9	61	0.217	13.99	2.45
5354	gestion	réseau		17	60	0.064	12.83	2.22

Index	N_1	N_2	R. expression	NC	LOG	FAG	IM3	h1
733	calcul	bilan		10	60	0.182	13.71	3.13
9435	sou-système	utilisateur	*sous-système utilisateur*	10	60	0.170	13.76	2.68
2616	appareil	mesure	*appareils de (mesure + mesures)*	7	60	0.268	14.82	0.94
4661	commutation	paquet		13	60	0.111	13.20	2.78
5488	architecture	réseau		14	60	0.054	12.84	1.31
336	signal	bande		38	59	0.076	13.17	3.95
6423	précision	pointage	*précision (de + du) pointage*	10	59	0.139	13.58	2.36
8736	communication	donnée	*(communication + communications) de données*	21	58	0.073	12.74	2.55
5220	procédure	coordination		11	58	0.134	13.34	2.94
7613	accroissement	température		8	58	0.182	14.10	1.51
5671	communication	satellite		34	57	0.052	12.89	2.55
1143	écoulement	trafic	*écoulement (du + de) trafic*	9	57	0.045	13.04	0.33
7707	sortie	amplificateur		15	57	0.095	12.81	3.30
1179	lancement	satellite		16	57	-0.001	12.12	0.78
8822	possibilité	correction	*(possibilité + possibilités) de correction*	10	56	0.143	13.43	2.91
4837	bit	information	*(bit + bits) d' information*	14	56	0.100	12.87	3.00
1665	échange	programme		8	56	0.166	13.95	1.19
6516	centre	exploitation		16	56	0.094	12.75	2.79
6897	distribution	télévision		13	56	0.085	12.85	2.50
3677	bruit	récepteur		14	55	0.111	12.72	3.39
6902	établissement	communication		13	55	0.074	12.76	2.35
486	signal	luminance		12	55	0.122	12.56	3.95
9352	traitement	donnée		19	54	0.066	12.52	2.25
2144	plupart	cas	*plupart des cas*	10	53	0.133	13.13	3.02
8601	suppression	écho	*(suppression + suppressions) d' écho*	9	53	0.112	13.28	1.75
8206	annuleurs	écho	*annuleurs d' écho*	8	53	0.111	13.47	0.67
8770	continuité	service	*continuité (de + du) service*	9	53	0.015	12.58	0.33
1566	orbite	satellite	*(orbite + orbites) des satellites*	21	52	0.023	12.22	1.22
6117	zone	service	*(zone + zones) de service*	16	52	0.060	12.44	2.04
339	refroidissement	air	*refroidissement (à + par) air*	6	52	0.312	14.58	2.09
7220	équipement	commutation	*(équipement + équipements) de commutation*	21	51	0.079	12.40	3.61
2732	cas	transmission		28	51	0.067	12.60	**4.12**
8866	qualité	transmission		21	51	0.061	12.40	3.16
8093	débit	symbole		8	51	0.177	13.37	2.61
2329	compensation	mouvement	*compensation (de + du) mouvement*	7	51	0.184	13.85	1.64
539	perte	gain		10	51	0.109	12.90	2.98
520	surface	terre	*surface de la terre*	12	51	0.041	12.40	1.65
2530	transport	message	*transport de (messages + message)*	6	50	0.157	13.99	0.41
5475	architecture	étoile	*architecture en étoile*	7	50	0.202	13.80	1.31
4444	signal	sortie		24	49	0.073	12.40	3.95
2851	station	type		21	49	0.075	12.33	3.51
355	période	éclipse		6	49	0.261	14.01	2.71
9291	technique	accès	*(technique + techniques) d' accès*	12	48	0.090	12.44	3.44
4206	communication	entreprise		10	48	0.120	12.63	2.55
9912	qualité	fonctionnement	*qualité de fonctionnement*	18	48	0.064	12.24	3.16
3509	ordinateur	serveur	*ordinateur serveur*	4	48	0.576	15.83	0.50
8921	caractéristique	qualité	*(caractéristique + caractéristiques) de qualité*	18	48	0.073	12.21	3.92
3225	intégration	service	*intégration (des + de) services*	10	47	0.026	12.27	1.40
6260	réseau	distribution	*(réseau + réseaux) de distribution*	13	47	0.096	12.11	2.94
7382	nombre	voie	*nombre (de + des) voies*	21	47	0.065	12.24	**3.91**
5421	calcul	rapport		12	47	0.077	12.33	3.13
2179	voisinage	saturation		7	47	0.146	13.40	2.08
4557	ambiguïté	phase	*ambiguïté de phase*	7	47	0.058	12.97	0.68
3096	espace	terre	*espace vers terre*	9	47	0.008	12.22	0.72
5507	type	limiteur	*type à limiteur*	9	46	0.121	12.17	**4.23**
3155	km	altitude	*km d' altitude*	4	46	0.507	15.56	0.50
1457	centre	commande	*centre de commande*	14	46	0.075	12.14	2.79
2919	emplacement	station		12	46	0.006	11.88	1.50
3837	saut	répéteur	*saut de répéteur*	9	46	0.045	12.41	1.27
9442	chaîne	réception		15	45	0.054	12.05	2.19

Index	N_1	N_2	R. expression	NC	LOG	FAG	IM3	h1
9433	objet	accusé		7	45	0.161	13.19	2.86
3648	puissance	bruit		25	45	0.062	12.27	2.99
1572	point	point	*point à point*	12	45	0.087	12.17	3.36
64	bit	échantillon	*bits par échantillon*	8	45	0.138	12.71	3.00
6478	convertisseur	fréquence		10	45	-0.014	11.79	0.69
3802	équipement	multiplication	*(équipement + équipements) de multiplication*	12	45	0.094	11.99	3.61
5757	fiche	notification	*(fiche + fiches) de notification*	4	45	0.453	15.34	0.50
18	ko	bit	*ko bits*	6	45	0.065	13.10	0.41
4044	antenne	réception		21	45	0.060	12.11	2.74
3781	transmission	programme		13	44	0.085	11.98	3.06
1779	bord	satellite		16	44	-0.001	11.63	1.50
6758	réseau	microstations	*(réseau + réseaux) de microstations*	13	44	0.087	11.90	2.94
4681	paquet	référence		11	44	0.068	12.15	2.65
5093	sortie	démodulateur	*sortie du démodulateur*	9	44	0.110	12.37	3.30
5052	commutation	bord		8	43	0.119	12.62	2.78
1106	lobe	antenne		10	43	-0.005	11.76	1.08
1268	bruit	brouillage		16	43	0.068	11.91	3.39
9161	numéro	séquence	*(numéro + numéros) de séquence*	5	43	0.216	14.02	1.15
6357	discrimination	polarisation		7	43	0.065	12.68	1.16
2702	bruit	quantification	*bruit de quantification*	10	43	0.101	12.02	3.39
5347	valeur	rapport		11	43	0.067	12.04	3.09
2330	alimentation	courant	*(alimentation + alimentations) en courant*	8	42	0.120	12.45	2.06
3572	station	réception		24	42	0.060	12.10	3.51
6237	service	télécommunication		23	42	0.059	12.04	3.23
4806	convertisseur-élévateur	fréquence	*convertisseur-élévateur de fréquence*	12	41	0.008	11.63	1.05
5299	réseau	télécommunication		26	41	0.061	12.14	2.94
4706	rapport	horizon	*rapport à l' horizon*	8	41	0.118	12.04	**3.38**
9095	objectif	disponibilité		7	41	0.130	12.66	1.06
898	différence	couleur	*différence de couleur*	5	41	0.231	13.73	2.37
7560	préambule	paquet		6	41	0.041	12.66	0.56
2416	prise	ressource	*prise de ressource*	5	41	0.208	13.74	1.73
2855	type	modulation		18	40	0.064	11.79	**4.23**
9174	dégradation	qualité		9	40	0.051	12.01	2.23
8144	câble	fibre	*(câble + câbles) à fibres*	4	40	0.358	14.66	0.97
6864	site	lancement	*site de lancement*	6	40	0.133	13.00	1.82
5561	augmentation	espacement	*augmentation de l' espacement*	5	40	0.225	13.38	3.16
7171	antenne	dimension		8	40	0.112	11.96	2.74
3235	linéariseur	type	*linéariseur du type*	7	40	0.027	12.19	1.12
8717	travail	génie	*travaux de génie*	4	39	0.342	14.43	1.89
5508	pays	membre	*pays membres + pays membre*	5	39	0.203	13.49	2.20
3619	détecteur	parole	*détecteur de parole*	6	39	0.112	12.81	1.85
8926	efficacité	utilisation	*efficacité d' utilisation*	6	39	0.115	12.81	2.31
9223	région	océan	*région de l' océan*	4	39	0.333	14.24	2.06
3822	sens	transmission	*sens de transmission*	10	39	0.013	11.61	1.87
854	circuit	référence		13	39	0.061	11.65	3.88
1474	centre	transit	*centre de transit*	8	39	0.106	11.99	2.79
1136	mode	fonctionnement	*(mode + modes) de fonctionnement*	15	38	0.048	11.60	3.01
4685	processus	application	*processus d' application*	7	38	0.101	12.26	3.12
9100	conception	système		14	38	0.022	11.47	2.42
5795	mode	multidestination	*mode multidestination*	6	38	0.142	12.26	3.01
1701	équipement	traitement		14	37	0.068	11.50	3.61
6936	point	saturation	*point de saturation*	9	37	0.083	11.68	3.36
4573	signalisation	ligne	*signalisation de ligne*	12	37	0.071	11.46	3.80
3432	récepteur	station		13	37	0.011	11.34	2.18
5644	boucle	réaction	*boucle de réaction*	5	36	0.171	12.98	2.04
4141	disposition	règlement	*(disposition + dispositions) du règlement*	5	36	0.175	12.87	3.19
5882	réflecteur	antenne		8	36	-0.035	11.24	0.89
6069	écran	visualisation	*écran de visualisation*	3	36	0.500	15.23	0.56
9096	système	poursuite	*(système + systèmes) de poursuite*	18	36	0.064	11.46	3.66
1763	cas	défaillance	*cas de défaillance*	10	36	0.078	11.40	**4.12**
9020	ligne	abonné		6	36	0.122	12.28	2.69

Index	N_1	N_2	R. expression	NC	LOG	FAG	IM3	h1
3152	trafic	répéteur		10	36	0.039	11.45	2.85
21	foi/fois	an	*fois par an*	4	36	0.254	13.77	2.03
9603	excursion	crête	*excursion de crête*	5	35	0.162	12.83	1.04
3629	niveau	bruit		19	35	0.047	11.55	3.60
9270	facteur	activité	*facteur d' activité*	6	35	0.124	12.11	2.50
3746	gain	multiplication		8	35	0.086	11.61	2.54
4962	sémaphore	message	*sémaphore de message*	6	35	0.085	12.21	2.17
2916	station	émission		21	35	0.052	11.61	3.51
1105	fonctionnement	mode	*fonctionnement en mode*	12	35	0.067	11.26	**4.82**
463	itinéraire	trafic	*(itinéraire + itinéraires) de trafic*	6	34	-0.025	11.61	0.74
4520	plan	fréquence		15	34	0.022	11.26	2.91
2901	émission	station		17	34	0.027	11.35	3.09
8205	système	commande		21	34	0.059	11.53	3.66
9670	entrée	amplificateur		10	34	0.051	11.33	2.93
968	combineurs	filtre	*combineurs à filtres*	6	34	0.061	12.00	1.85
7685	maintien	poste	*maintien à poste*	5	34	0.137	12.53	1.35
1768	défaillance	équipement		8	34	0.028	11.42	2.64
3639	appel	offre	*(appel + appels) d' offres*	4	34	0.218	13.02	3.28
5377	réseau	étoile		9	33	0.078	11.13	2.94
2142	loi	illumination	*loi d' illumination*	3	33	0.408	14.64	1.01
4596	publication	renseignement		3	33	0.362	14.64	0.56
820	code	rendement		6	33	0.113	11.78	2.71
8930	tonalité	essai	*tonalité d' essai*	4	33	0.133	13.20	0.96
916	fréquence	échantillonnage	*fréquence d' échantillonnage*	9	33	0.072	11.23	3.90
684	choix	emplacement		5	33	0.142	12.26	3.10
5568	énergie	bit		7	33	0.050	11.57	2.62
2694	message	résultat	*message de résultat*	4	33	0.207	13.08	2.13
6318	récepteur	poursuite	*récepteur de poursuite*	7	33	0.055	11.55	2.18
9350	accès	assignation	*accès multiple avec assignation*	7	33	0.084	11.53	2.66
1262	volume	trafic	*volume de trafic*	6	33	-0.024	11.44	1.00
8236	cadre	système		11	33	0.002	11.01	2.13
1798	méthode	calcul	*(méthode + méthodes) de calcul*	7	32	0.085	11.48	3.04
3518	détecteur	mouvement	*détecteur de mouvement*	5	32	0.094	12.31	1.85
5511	emploi	technique		6	32	0.086	11.80	3.27
3368	détection	erreur		8	32	0.008	11.17	1.97
6285	modulation	impulsion	*modulation par impulsions*	8	32	0.072	11.25	2.52
1833	pompe	chaleur	*(pompe + pompes) à chaleur*	3	32	0.280	14.23	0.56
4941	corrélation	erreur	*corrélation des erreurs*	6	32	-0.032	11.28	0.85
2995	alimentation	antenne		15	31	0.024	11.10	2.06
2986	affaiblissement	transmission	*affaiblissement de transmission*	12	31	0.023	11.02	2.45
7169	dimension	antenne		9	31	-0.015	10.89	1.94
1587	mise	point	*mise au point*	9	31	0.057	11.09	2.16
5834	transmission	satellite		39	31	0.044	12.32	3.06
1745	signal	image		12	31	0.058	10.95	3.95
4910	répartition	code	*répartition en code*	7	31	0.047	11.30	1.48

Indefinite values of log-likehood coefficient

Index	N_1	N_2	R. expression	NC	LOG	FAG	IM3	h1
4551	abaisseur	fréquence	*(abaisseur + abaisseurs) de fréquence*	12	∞	0.011	12.69	0.00
5556	adaptateur	interface	*adaptateur d' interfaces*	2	∞	-0.175	10.49	0.00
3098	adjudication	contrat	*adjudication du contrat*	2	∞	0.354	14.47	0.00
9240	agilité	fréquence	*agilité (en + de) fréquence*	7	∞	-0.070	11.13	0.00
1466	aiguille	montre	*aiguilles d' une montre*	2	∞	0.646	15.47	0.00
5358	allumage	moteur	*allumage (du moteur + des moteurs)*	5	∞	0.207	14.36	0.00
4942	amortissement	nutation	*amortissement de la nutation*	2	∞	0.463	14.89	0.00
8374	angle	obliquité	*angle d' obliquité*	2	∞	0.116	10.52	1.69
9037	annulation	écho	*annulation d' écho*	3	∞	-0.094	11.34	0.00
8388	annuleur	écho	*annuleur d' écho*	5	∞	0.028	12.81	0.00
3052	antenne	tore	*(antenne + antennes) tore*	2	∞	0.065	8.86	2.74
996	arséniure	gallium	*arséniure de gallium*	11	∞	0.849	17.93	0.00
502	asservissement	antenne	*asservissement de l' antenne*	2	∞	-0.286	7.70	0.00
2272	atop	linéariseur	*atop avec linéariseur*	5	∞	0.567	16.11	0.00
4228	attente	numérotation	*attente après numérotation*	2	∞	-0.020	12.30	0.00
771	axe	déclinaison	*axe de déclinaison*	2	∞	0.148	11.22	2.40
9723	banque	donnée	*banques de données*	3	∞	-0.184	9.54	0.00
171	base	tarif	*base des tarifs*	2	∞	0.107	10.28	3.84
2140	bas	page	*bas de page*	2	∞	0.457	14.47	1.04
7986	bit	parité	*(bit + bits) de parité*	3	∞	0.132	11.20	3.00
4288	bouton	poussoir	*bouton poussoir*	2	∞	0.373	13.89	1.56
133	câblage	baie	*câblage entre baies*	2	∞	0.457	14.47	1.04
5029	caractérisation	information	*caractérisation de (ces informations + l' information)*	2	∞	-0.209	9.89	0.00
8115	carte	crédit	*cartes de crédit*	2	∞	0.373	13.89	1.56
2658	centrage	spectre	*centrage du spectre*	2	∞	-0.145	10.95	0.00
1476	centre	télémaintenance	*centre de télémaintenance*	2	∞	0.080	9.45	2.79
431	chemin	roulement	*chemin de roulement*	2	∞	0.646	15.47	0.00
550	circuit	silencieux	*circuit de silencieux*	2	∞	0.079	9.42	3.88
4202	circulateur	ferrite	*circulateur à ferrite*	2	∞	0.528	14.89	0.64
3562	codage	voix	*codage de la voix*	2	∞	0.120	10.61	3.34
792	codeur-décodeur	récurrence	*codeur-décodeur à récurrence*	2	∞	0.108	10.30	2.31
9107	collège	formation	*(collège + collèges) de formation*	2	∞	0.181	13.66	0.00
5352	collecteur	dépression	*collecteur (en + à) dépression*	3	∞	0.436	14.64	1.67
7378	commission	étude	*commission d' études*	6	∞	0.358	15.39	0.00
2589	compresseur	émission	*compresseur à l' émission*	2	∞	-0.260	8.63	0.00
7997	conductivité	sol	*conductivité du sol*	2	∞	-0.059	11.95	0.00
2853	conférence	radiocommunication	*conférence administrative (extraordinaire + mondiale) des radiocommunications*	4	∞	-0.014	12.30	0.00
4894	conseil	gouverneur	*conseil des gouverneurs*	2	∞	0.409	14.15	1.05
4787	convention	télécommunication	*convention internationale des télécommunications*	2	∞	-0.264	8.50	0.00
8051	côte	côte	*côte à côte*	2	∞	0.646	15.47	0.00
7557	cristal	glace	*cristaux de glace*	2	∞	0.646	15.47	0.00
1493	croissance	trafic	*croissance du trafic*	3	∞	-0.165	10.03	0.00
1187	débordement	mémoire	*débordement de la mémoire*	2	∞	-0.029	12.22	0.00
3540	découplage	polarisation	*découplage (des polarisations + de polarisation)*	2	∞	-0.205	9.96	0.00
2984	déformation	antenne	*(déformations + déformation) de l' antenne*	2	∞	-0.286	7.70	0.00
9191	démultiplexeur	entrée	*démultiplexeur d' entrée*	5	∞	-0.049	11.76	0.00
1658	déploiement	satellite	*déploiement du satellite*	2	∞	-0.310	6.44	0.00
2723	db	octave	*db par octave*	2	∞	0.152	11.30	2.94

Index	N_1	N_2	R. expression	NC	LOG	FAG	IM3	h1
3825	dent	scie	*dent de scie*	2	∞	0.646	15.47	0.00
724	diode	laser		2	∞	0.289	13.15	2.03
1732	diode	tunnel	*diodes tunnel*	2	∞	0.289	13.15	2.03
6655	dissipation	chaleur		2	∞	0.146	13.47	0.00
1157	dizaine	mbit	*dizaines de mbit*	2	∞	0.373	13.89	1.33
197	dizaine	millisecondes	*dizaines de millisecondes*	2	∞	0.373	13.89	1.33
8266	durée	reconnaissance	*(durée + durées) de reconnaissance*	2	∞	0.122	10.66	2.85
4369	effet	masque	*effet de masque*	3	∞	0.146	11.49	3.58
4196	égalisation	temps		11	∞	0.150	14.46	0.00
9545	égalité	droit	*égalité (de + des) droits*	5	∞	0.776	16.79	0.00
5626	élément	contact	*élément de contact*	2	∞	0.141	11.08	2.83
5111	élévation	température		3	∞	-0.028	12.18	0.00
8932	embrouillage	donnée	*embrouillage de données*	2	∞	-0.268	8.37	0.00
8532	entité	propriétaire	*entité propriétaire*	2	∞	0.354	14.47	0.00
5581	énergie	désembrouillage	*énergie par désembrouillage*	2	∞	0.162	11.47	2.62
1663	équinoxe	automne	*équinoxe d' automne*	3	∞	0.551	15.32	0.67
1542	équinoxe	printemps	*équinoxe de printemps*	2	∞	0.409	14.15	0.67
6381	évanouissement	propagation	*(évanouissement + évanouissements) de propagation*	5	∞	-0.039	11.92	0.00
6621	extenseur	réception	*extenseur à la réception*	2	∞	-0.263	8.54	0.00
813	fibre	carbone	*fibres de carbone*	2	∞	0.528	14.89	0.64
6543	filtre	miroir	*filtres miroirs*	2	∞	0.141	11.08	3.09
6429	focalisation	bobine	*focalisation (à bobine + par bobines)*	2	∞	0.457	14.47	0.69
2401	fonction	tri	*fonction de tri*	2	∞	0.065	8.83	4.52
9193	formatage	donnée		2	∞	-0.268	8.37	0.00
1897	glossaire	fascicule	*glossaire du fascicule*	2	∞	0.646	15.47	0.00
2779	indice	modulation	*indice de modulation*	4	∞	-0.071	11.51	0.00
8740	interconnectivité	station	*interconnectivité entre stations*	3	∞	-0.213	8.62	0.00
708	interface	saisie	*interface de commande et de saisie*	2	∞	0.120	10.61	2.40
6521	interliaison	réseau	*interliaison de réseaux*	2	∞	-0.275	8.12	0.00
3298	intertrame	mouvement	*intertrame à mouvement*	2	∞	-0.100	11.52	0.00
5858	inversion	phase	*inversion de phase*	5	∞	-0.008	12.36	0.00
7232	jonction	orthomode	*(jonction + jonctions) orthomode*	3	∞	0.418	15.06	0.00
7564	juridiction	partie	*juridiction d' une partie*	3	∞	0.174	13.83	0.00
1077	laps	temps	*laps de temps*	2	∞	-0.226	9.54	0.00
8406	largeur	fenêtre	*largeur de la fenêtre*	2	∞	0.059	8.56	0.38
9062	levée	ambiguïté		4	∞	0.644	16.15	0.00
897	lien	codage	*liens entre codage*	2	∞	-0.197	10.11	0.00
8145	ligne	flèche	*lignes à une flèche*	2	∞	0.127	10.77	2.69
5956	luminance	amplitude	*luminance à l' amplitude*	2	∞	-0.152	10.86	0.00
7822	mécanisme	captage	*(mécanisme + mécanismes) de captage*	2	∞	0.236	12.56	2.21
7264	méthode	évaluation	*méthodes d' évaluation*	2	∞	0.089	9.76	3.04
7287	méthode	détermination	*(méthode + méthodes) de détermination*	3	∞	0.120	10.93	3.04
3045	matériel	démonstration		2	∞	0.457	14.47	1.04
1510	maximum	vraisemblance	*maximum de vraisemblance*	5	∞	0.481	15.41	1.73
9119	modèle	fiche	*modèles de fiches*	2	∞	0.204	12.15	1.37
5809	modulation	inversion	*(modulation + modulations) par inversion*	5	∞	0.151	12.06	2.52
2787	modulation	multiphase	*modulation multiphase*	2	∞	0.079	9.42	2.52
4512	mono-impulsion	multimode	*mono-impulsion multimode*	2	∞	0.457	14.47	0.69
3259	monture	antenne	*(monture + montures) d' antenne*	2	∞	-0.286	7.70	0.00
9341	moteur	périgée	*moteur de périgée*	2	∞	0.143	11.11	1.22
1827	moyen	navette	*moyen de la navette*	2	∞	0.075	9.27	4.24
6488	multiplexeur	sortie	*multiplexeur de sortie*	5	∞	-0.052	11.71	0.00
789	orbite	parking	*orbite de parking*	2	∞	0.141	11.08	1.22
2447	ordre	grandeur	*ordre de grandeur*	3	∞	0.251	13.06	2.61
4709	ouverture	mi-puissance	*ouverture à mi-puissance*	2	∞	0.254	12.77	1.38
7299	personne	juridiction	*personne sous juridiction*	2	∞	0.646	15.47	0.00
8856	pièce	rechange	*pièces de rechange*	3	∞	0.366	14.83	0.00
3613	plan	équateur	*plan de l' équateur*	3	∞	0.155	11.66	2.91
4542	plan	allotissement	*plan d' allotissement*	2	∞	0.115	10.49	2.91
2334	plan	découpage	*plan de découpage*	2	∞	0.115	10.49	2.91

Index	N_1	N_2	R. expression	NC	LOG	FAG	IM3	h1
546	plaque	diélectrique	*plaque diélectrique*	6	∞	0.616	16.32	0.90
3086	plaque	quart	*plaque quart*	3	∞	0.390	14.32	0.90
6938	point	interception	*point d' interception*	7	∞	0.186	13.03	3.36
7067	polariseur	onde	*polariseur quart d' onde*	2	∞	-0.204	10.00	0.00
4530	pompage	fréquence		2	∞	-0.290	7.52	0.00
5261	porteur	plage	*porteuse par plage*	2	∞	0.088	9.70	3.35
2475	pression	rayonnement	*pression du rayonnement*	2	∞	-0.171	10.56	0.00
4684	processus	interrogation	*processus d' interrogation*	2	∞	0.138	11.01	3.12
829	profondeur	décharge	*profondeur de décharge*	2	∞	0.528	14.89	0.64
9897	pureté	polarisation	*pureté de polarisation*	7	∞	0.088	13.58	0.00
5914	quadrature	phase	*quadrature de phase*	3	∞	-0.122	10.89	0.00
7338	quart	onde	*quart d' onde*	3	∞	-0.105	11.17	0.00
4961	rail	azimut	*rail d' azimut*	2	∞	0.073	13.01	0.00
2157	rapport	découplage	*rapport de découplage*	2	∞	0.063	8.74	3.38
6448	rapport	isolement	*rapport d' isolement*	2	∞	0.063	8.74	3.38
1009	récurrence	redondance	*récurrence avec redondance*	2	∞	-0.045	12.08	0.00
518	refroidissement	hélium	*refroidissement (à + par) hélium*	2	∞	0.204	12.15	2.09
815	rendement	combustible	*rendement du combustible*	2	∞	0.143	11.11	2.75
2102	repliement	spectre	*repliement du spectre*	2	∞	-0.145	10.95	0.00
8545	reportage	actualité		3	∞	0.711	16.06	0.00
5026	réorganisation	acheminement	*réorganisation de l' acheminement*	2	∞	0.012	12.56	0.00
7064	répéteur	écrêteur	*répéteurs à écrêteur*	2	∞	0.088	9.70	1.44
6355	réseau	collecte	*(réseau + réseaux) de collecte*	3	∞	0.066	9.18	2.94
6277	réussite	lancement		2	∞	-0.095	11.56	0.00
4696	réutilisation	fréquence		27	∞	0.137	15.03	0.00
4384	risque	accident	*risques d' accident*	2	∞	0.289	13.15	1.47
3778	rotation	faraday	*rotation de faraday*	2	∞	0.204	12.15	2.51
9004	saisie	donnée	*saisie de données*	2	∞	-0.268	8.37	0.00
4855	saut	cycle	*(saut + sauts) de cycle*	2	∞	0.215	12.30	1.27
3516	serveur	station	*serveur de la station*	2	∞	-0.292	7.45	0.00
3990	service	radiocommunication	*(service(s) de radiocommunication) + (services de radiocommunications)*	10	∞	0.164	13.08	3.23
5832	signalisation	décade	*signalisation à décade*	3	∞	0.075	9.57	3.80
3928	silence	conversation	*silences de la conversation*	2	∞	-0.165	10.66	0.00
1760	socle	antenne	*socle de l' antenne*	3	∞	-0.206	8.87	0.00
3680	sortie	multiplicateur	*sortie du multiplicateur*	2	∞	0.092	9.86	3.30
9395	sou-réseau	accès	*sous-réseau d' accès*	2	∞	-0.195	10.15	0.00
3878	station	correspondance	*stations en correspondance*	3	∞	0.072	9.44	3.51
8600	suppresseur	écho	*suppresseur d' écho*	13	∞	0.267	15.57	0.00
7083	synthétiseur	fréquence		12	∞	0.011	12.69	0.00
8246	système	avertissement	*(système + systèmes) d' avertissement*	2	∞	0.035	7.08	3.66
194	télésurveillance	alarme	*télésurveillance des alarmes*	2	∞	-0.065	11.89	0.00
2053	technologie	invar	*technologie de l' invar*	2	∞	0.254	12.77	2.35
212	tec	arséniure	*tec à arséniure*	3	∞	0.486	15.32	0.00
5118	température	brillance	*température de brillance*	2	∞	0.077	9.32	0.65
8700	temps	montée	*temps de montée*	2	∞	0.077	9.33	1.80
8056	théorème	réciprocité	*théorème de réciprocité*	2	∞	0.457	14.47	0.69
401	trajet	descendant	*trajets descendants*	2	∞	0.155	11.34	2.74
3700	transistor	effet	*(transistor + transistors) à effet*	10	∞	0.516	16.66	0.00
2119	transit	satellite	*transit à satellites*	6	∞	-0.128	9.61	0.00
192	tri	cliques/clique	*tri (des + de) cliques*	2	∞	0.646	15.47	0.00
6019	type	famille	*types de famille*	2	∞	0.053	8.25	4.23
1484	union	république	*union des républiques*	2	∞	0.409	14.15	1.05
1499	véhicule	lancement	*(véhicule + véhicules) de lancement*	3	∞	0.028	12.73	0.00
5081	vérification	carte	*vérification (de + des) cartes*	2	∞	0.305	13.30	1.89
8884	visée	antenne	*visée de l' antenne*	2	∞	-0.286	7.70	0.00
7187	voie	conséquence	*voie de conséquence*	2	∞	0.065	8.83	3.68
6528	volant	inertie	*volant d' inertie*	3	∞	0.418	15.06	0.00
25	vol	bit	*vol de bits*	2	∞	-0.195	10.15	0.00
7451	zone	ouest	*zone ouest*	2	∞	0.116	10.52	2.04

N ADJ

Sorting of log-likehood coefficient

Index	N	Adj	R. expression	NC	LOG	FAG	IM3	h2
1708	station	terrien	*stations terriennes + station terrienne*	750	2933	0.842	22.47	0.24
3680	débit	binaire	*débit binaire + débits binaires*	134	715	0.675	19.45	1.24
4116	accès	multiple	*accès multiple*	105	605	0.664	19.09	1.74
3588	voie	téléphonique		118	511	0.512	18.52	2.24
2847	liaison	montant	*liaison montante + liaisons montantes*	88	456	0.519	18.09	0.26
1282	liaison	descendant		77	407	0.491	17.75	0.14
1580	secteur	spatial	*secteur spatial*	79	341	0.409	17.41	2.21
1795	service	fixe	*service fixe + services fixes*	66	326	0.467	17.42	1.19
3836	lobe	latéral	*lobes latéraux*	40	299	0.699	17.93	0.72
1952	faisceau	hertzien	*faisceau hertzien + faisceaux hertziens*	35	244	0.586	17.22	0.40
1157	puissance	surfacique	*puissance surfacique + puissances surfaciques*	35	231	0.502	16.76	0.13
2013	polarisation	circulaire		36	204	0.443	16.49	1.15
3106	oscillateur	local	*oscillateurs locaux + oscillateur local*	32	200	0.434	16.45	2.94
1515	bruit	thermique		35	183	0.393	16.15	1.92
2936	polarisation	rectiligne	*polarisation rectiligne + polarisations rectilignes*	28	181	0.450	16.16	0.29
1499	erreur	binaire	*erreur binaire + erreurs binaires*	40	158	0.263	15.40	1.24
2207	espace	libre	*espace libre*	18	156	0.678	16.83	0.63
1364	groupe	primaire		27	152	0.385	15.82	2.00
3546	amplificateur	paramétrique	*amplificateur paramétrique + amplificateurs paramétriques*	19	142	0.534	16.22	1.05
1022	signal	vocal	*signal vocal + signaux vocaux*	32	140	0.286	15.09	1.41
188	commande	automatique	*commande automatique*	20	137	0.421	15.80	2.48
3082	réflecteur	secondaire	*réflecteur secondaire + réflecteurs secondaires*	24	135	0.360	15.51	1.50
2810	décision	souple		15	131	0.642	16.44	0.73
5	orbite	géostationnaire		18	123	0.442	15.70	0.69
3261	démodulation	cohérent		15	118	0.521	15.94	1.44
939	équipement	terminal	*équipements terminaux + équipement terminal*	19	117	0.361	15.06	0.70
4255	atmosphère	clair	*atmosphère claire*	12	116	0.693	16.42	0.56
1193	groupe	secondaire		22	114	0.304	14.95	1.50
1794	courant	continu	*courant continu*	18	114	0.350	15.22	2.17
889	filtre	passe-bande	*filtre passe-bande + filtres passe-bande*	15	113	0.455	15.41	0.23
3736	système	national		39	111	0.197	14.43	2.55
3160	polarisation	orthogonal	*polarisations orthogonales + polarisation orthogonale*	19	109	0.322	14.75	0.87
84	circuit	fictif	*circuit fictif + circuits fictifs*	20	108	0.303	14.64	0.78
3466	onde	progressif	*ondes progressives*	19	104	0.314	14.81	1.68
2606	section	spécial	*sections spéciales + section spéciale*	14	101	0.415	15.31	2.32
3615	réflecteur	principal	*réflecteur principal*	22	100	0.235	14.40	3.20
2087	transmission	numérique		38	97	0.151	14.01	3.66
1586	station	distant	*stations distantes + station distante*	44	96	0.160	13.80	1.35
700	fréquence	intermédiaire	*fréquence intermédiaire + fréquences intermédiaires*	21	96	0.243	14.10	1.48
678	exemple	typique	*exemples typiques + exemple typique*	17	95	0.286	14.57	2.54
1581	destination	multiple	*destinations multiples + destination multiple*	20	91	0.178	13.90	1.74
2246	utilisateur	final	*utilisateurs finals + utilisateur final*	12	88	0.415	15.08	1.84
1780	codage	correcteur	*codage correcteur + codages correcteurs*	16	86	0.272	14.26	1.63
877	refroidissement	thermoélectrique	*refroidissement thermoélectrique*	9	85	0.606	15.64	0.33
3135	organisation	international	*organisations internationales + organisation internationale*	18	84	0.180	13.79	3.28

Index	N	Adj	R. expression	NC	LOG	FAG	IM3	h2
1662	heure	actuel	*heure actuelle*	12	80	0.304	14.46	2.77
2040	code	correcteur	*code correcteur + codes correcteurs*	16	78	0.237	13.87	1.63
3610	règle	général	*règle générale*	14	78	0.162	13.57	3.58
3518	énergie	électrique	*énergie électrique*	14	78	0.257	14.06	2.44
679	satellite	géostationnaire		18	78	0.211	13.56	0.69
4372	densité	spectral	*densité spectrale + densités spectrales*	11	76	0.317	14.44	2.00
315	affaiblissement	atmosphérique	*affaiblissements atmosphériques + affaiblissement atmosphérique*	10	76	0.385	14.74	1.71
2583	fréquence	porteur	*fréquences porteuses + fréquence porteuse*	17	73	0.201	13.33	1.44
4172	compatibilité	technique	*compatibilité technique*	13	72	0.196	13.67	3.00
634	distance	minimal	*distance minimale*	12	71	0.264	13.96	2.50
2724	mode	continu		14	70	0.207	13.58	2.17
2706	bande	étroit		15	67	0.201	13.23	1.12
3842	état	solide	*état solide*	8	66	0.444	14.76	1.07
269	conduit	numérique	*conduit numérique + conduits numériques*	15	66	0.066	12.49	3.66
1409	émission	brouilleuse	*émission brouilleuse*	8	65	0.416	14.66	0.88
4128	câble	coaxial	*câble coaxial + câbles coaxiaux*	8	65	0.426	14.61	0.82
1130	circuit	téléphonique	*circuits téléphoniques + circuit téléphonique*	29	65	0.120	13.03	2.24
2639	transmission	analogique		20	65	0.144	12.99	3.16
2860	position	orbital	*position orbitale + positions orbitales*	9	64	0.325	14.17	1.55
896	gain	nominal	*gain nominal*	11	63	0.244	13.61	2.16
2628	limiteur	progressif	*limiteur progressif*	9	63	0.257	13.90	1.68
931	élément	rayonnant	*élément rayonnant + éléments rayonnants*	9	63	0.316	13.80	0.60
2156	polarisation	elliptique	*polarisation elliptique*	11	59	0.220	13.03	0.95
2053	courant	alternatif	*courant alternatif*	7	57	0.387	14.08	0.38
2660	faisceau	étroit		12	57	0.190	12.92	1.12
4063	qualité	subjectif		9	56	0.256	13.42	1.87
161	surface	équivalent	*surface équivalente + surfaces équivalentes*	8	55	0.284	13.73	1.91
2921	réseau	public		12	55	0.177	12.51	0.79
1595	station	central		40	55	0.118	12.89	2.05
1083	panneau	solaire	*panneaux solaires + panneau solaire*	8	54	0.244	13.56	2.36
1132	moment	cinétique	*moment cinétique*	5	54	0.629	15.16	0.45
3531	transmission	télévisuel	*transmissions télévisuelles + transmission télévisuelle*	11	54	0.193	12.77	1.38
2139	enveloppe	constant	*enveloppe constante*	6	53	0.378	14.31	2.01
1847	niveau	admissible	*niveaux admissibles + niveau admissible*	10	51	0.197	12.68	1.27
4083	signal	vidéo	*signaux vidéo + signal vidéo*	14	51	0.148	12.28	2.12
1579	porteur	multiple	*porteuses multiples + porteuse multiple*	20	51	0.104	12.35	1.74
1711	question	relatif	*questions relatives*	9	51	0.129	12.73	**3.43**
3007	poursuite	automatique	*poursuite automatique*	8	51	0.163	13.02	2.48
1009	onde	triangulaire	*onde triangulaire*	8	51	0.247	13.00	0.76
2120	charge	essentiel		9	50	0.195	12.92	2.49
2471	algorithme	probabiliste	*algorithmes probabilistes + algorithme probabiliste*	6	49	0.373	13.99	0.90
4465	efficacité	spectral	*efficacité spectrale*	7	49	0.206	13.25	2.00
4361	schéma	fonctionnel	*schéma fonctionnel + schémas fonctionnels*	8	49	0.224	13.05	1.96
3334	antenne	isotrope	*antenne isotrope + antennes isotropes*	8	48	0.232	12.91	0.88
3583	ouverture	angulaire	*ouverture angulaire*	8	48	0.186	12.95	2.37
3863	système	mondial	*système mondial + systèmes mondiaux*	14	48	0.136	12.08	2.09
1980	signal	transmis		12	48	0.148	12.06	1.36
2191	plan	équatorial	*plan équatorial*	6	47	0.344	13.70	1.00
1641	radiocommunication	spatial	*radiocommunication spatiale*	11	47	0.045	11.85	2.21
3922	signal	reçu	*signal reçu*	10	47	0.163	12.03	0.79
3062	décodage	séquentiel	*décodage séquentiel*	6	47	0.335	13.63	1.00
3507	dimension	moyen	*dimensions moyennes + dimension moyenne*	8	46	0.134	12.62	3.05

Index	N	Adj	R. expression	NC	LOG	FAG	IM3	h2
2687	assignation	préalable	*assignation préalable*	5	45	0.412	14.16	1.15
2136	occupation	spectral		6	45	0.184	13.10	2.00
1796	service	mobile	*service mobile + services mobiles*	12	44	0.136	11.86	1.84
2627	fréquence	central	*fréquence centrale + fréquences centrales*	18	44	0.106	11.96	2.05
3193	longueur	variable		6	44	0.254	13.34	2.32
1011	espacement	angulaire	*espacements angulaires + espacement angulaire*	9	44	0.152	12.34	2.37
4157	polynôme	générateur	*polynôme générateur*	4	44	0.576	14.72	0.50
2722	réseau	national		21	43	0.097	11.99	2.55
745	caractéristique	technique	*caractéristiques techniques*	14	43	0.106	11.88	3.00
2712	segment	spatial	*segment spatial*	10	42	0.029	11.54	2.21
3360	formule	suivant		10	42	0.087	11.91	**3.80**
1049	onde	lent	*ondes lentes*	7	42	0.208	12.42	1.03
3694	bande	latéral	*bande latérale + bandes latérales*	13	42	0.115	11.79	0.72
3289	gestion	opérationnel	*gestion opérationnelle*	6	41	0.202	12.89	2.68
458	fréquence	supérieur	*fréquences supérieures + fréquence supérieure*	14	41	0.110	11.66	2.99
4125	système	existant	*systèmes existants + système existant*	11	40	0.126	11.55	1.69
2720	régulation	thermique	*régulation thermique*	7	40	0.108	12.20	1.92
2943	réseau	terrestre	*réseaux terrestres + réseau terrestre*	12	40	0.118	11.60	2.40
1697	charge	utile	*charge utile + charges utiles*	7	40	0.175	12.35	1.95
1284	onde	radioélectrique	*onde radioélectrique + ondes radioélectriques*	11	40	0.105	11.70	2.64
2891	détection	cohérent	*détection cohérente*	5	39	0.189	12.99	1.44
2192	rapport	axial	*rapport axial*	5	39	0.303	13.13	0.60
724	valeur	typique	*valeur typique + valeurs typiques*	12	39	0.107	11.62	2.54
2434	température	ambiant	*température ambiante + températures ambiantes*	5	39	0.291	12.94	0.45
397	mot	unique	*mot unique + mots uniques*	7	39	0.122	12.18	3.09
3864	république	fédéral	*république fédérale*	5	39	0.286	12.89	0.45
882	république	socialiste	*république socialiste + républiques socialistes*	5	39	0.286	12.89	0.45
1700	alimentation	périscopique	*alimentation périscopique*	6	39	0.219	12.39	1.00
2488	température	physique	*température physique*	7	39	0.160	12.22	2.34
1280	traitement	numérique	*traitement numérique*	14	38	0.040	11.34	3.66
328	cas	particulier	*cas particulier + cas particuliers*	9	38	0.106	11.76	**3.47**
2058	téléphonie	rural		6	38	0.189	12.57	1.84
3375	représentation	schématique	*représentation schématique*	4	38	0.430	13.87	0.50
1549	alimentation	électrique	*alimentation électrique*	9	38	0.119	11.76	2.44
680	satellite	national		18	38	0.086	11.61	2.55
1296	formule	approximatif	*formule approximative*	5	37	0.264	12.80	0.90

Indefinite values of log-likehood coefficient

Index	N	Adj	R. expression	NC	LOG	FAG	IM3	h2
2178	absorption	atmosphérique	*absorption atmosphérique*	5	∞	0.233	13.42	1.71
1927	accroissement	apparent	*accroissement apparent*	2	∞	0.279	13.04	1.33
2923	administration	notificatrice	*administration notificatrice*	3	∞	0.390	13.21	0.00
925	agent	local	*agents locaux + agent local*	7	∞	0.070	12.28	2.94
323	ailette	métallique	*ailettes métalliques + ailette métallique*	2	∞	0.354	13.36	1.04
922	aimant	permanent	*aimant permanent + aimants permanents*	7	∞	0.351	14.39	2.36
1560	alignement	plésiochrone	*alignement plésiochrone*	2	∞	0.146	12.36	1.73
3003	alimentation	frontal	*alimentation frontale*	6	∞	0.284	12.98	0.00
1625	alimentation	ininterruptible	*alimentation ininterruptible*	9	∞	0.365	14.15	0.00
3199	amplificateur	excitateur	*amplificateur excitateur*	3	∞	0.208	11.40	0.00
1029	antenne	imparfait		2	∞	0.129	9.72	0.00
2208	article	manquant	*articles manquants*	2	∞	0.457	13.36	0.00
4353	assemblée	plénier	*assemblée plénière*	2	∞	0.646	14.36	0.00
1516	atop	seul	*atop seul*	4	∞	0.566	14.78	0.87
2197	attention	particulier	*attention particulière*	3	∞	-0.074	10.51	3.47
3827	autorité	responsable	*autorités responsables*	2	∞	0.224	12.78	1.56
3668	béton	armé	*béton armé*	2	∞	0.646	14.36	0.00
4059	bande	appariées	*bandes appariées*	3	∞	0.113	9.64	0.00
2410	bande	interdit	*bandes interdites*	2	∞	0.084	8.47	0.00
1490	bande	passant	*bande passante + bandes passantes*	10	∞	0.244	13.11	0.00
41	bobine	électromagnétique	*bobines électromagnétiques + bobine électromagnétique*	4	∞	0.417	14.19	0.69
2401	brouillage	potentiel	*brouillages potentiels + brouillage potentiel*	2	∞	0.152	10.19	0.00
2353	bruit	cosmique	*bruit cosmique*	2	∞	0.093	8.76	0.00
2154	bruit	impulsif	*bruit impulsif*	9	∞	0.254	13.10	0.00
3912	câble	sous-marins	*câbles sous-marins*	4	∞	0.354	13.19	0.00
3698	câble	sous-marin	*câble sous-marin*	5	∞	0.409	13.84	0.00
3657	cavité	couplées	*cavités couplées*	2	∞	0.409	13.04	0.00
754	cellule	solaire	*cellules solaires*	5	∞	0.172	13.01	2.36
3580	circuit	interrupteur	*circuit interrupteur*	2	∞	0.083	8.43	0.00
220	circulaire	hebdomadaire	*circulaire hebdomadaire*	5	∞	0.614	15.01	0.00
4234	coût	modique	*coût modique*	2	∞	0.179	10.66	0.00
140	codage	adaptatif		2	∞	0.128	9.69	0.00
472	code	auto-orthogonaux	*codes auto-orthogonaux*	5	∞	0.212	11.94	0.00
4084	code	poinçonnés		3	∞	0.151	10.47	0.00
2334	combineurs	hybride	*combineurs hybrides*	2	∞	0.181	12.56	1.75
1713	compresseur-extenseur	syllabique	*compresseur-extenseur syllabique*	2	∞	0.094	12.04	1.47
2776	compte	tenu	*compte tenu*	23	∞	0.896	17.89	0.00
2046	concentration	numérique	*concentration numérique*	34	∞	0.211	14.95	3.66
560	conférence	administratif	*conférence administrative*	4	∞	0.353	13.90	1.77
2939	consultation	officiel	*consultations officielles*	2	∞	0.346	12.56	0.00
3192	consultation	officieux	*consultations officieuses*	2	∞	0.346	12.56	0.00
3991	cornet	cannelé	*cornet cannelé*	2	∞	0.254	11.66	0.00
1672	couple	perturbateur	*couple perturbateur + couples perturbateurs*	7	∞	0.715	15.81	0.00
3376	courrier	électronique	*courrier électronique*	3	∞	0.144	12.53	1.89
1688	couverture	hémisphérique		6	∞	0.382	13.83	0.00
3369	couverture	zonale		5	∞	0.340	13.31	0.00
2821	décision	ferme	*décisions fermes*	2	∞	0.199	10.97	0.00
3595	déflection	mécanique	*déflection mécanique*	2	∞	-0.091	10.50	2.98
4325	description	général	*description générale*	4	∞	-0.086	10.14	3.58
846	diaphonie	intelligible	*diaphonie intelligible*	2	∞	0.646	14.36	0.00
4201	directeur	général	*directeur général*	3	∞	-0.147	9.31	3.58
4071	disponibilité	accrue	*disponibilité accrue*	2	∞	0.276	11.90	0.00

Index	N	Adj	R. expression	NC	LOG	FAG	IM3	h2
2065	dissipation	thermique	*dissipation thermique*	2	∞	-0.177	9.36	1.92
2234	distorsion	intersymbole	*distorsion intersymbole*	6	∞	0.737	15.73	0.00
2400	domaine	fréquentiel	*domaine fréquentiel*	2	∞	0.289	12.04	0.00
1061	éclaircissement	nécessaire	*éclaircissements nécessaires*	2	∞	-0.185	9.23	**3.82**
2430	égaliseur	accordable		2	∞	0.528	13.78	0.00
849	embrouilleur	pseudo-aléatoire	*embrouilleur pseudo-aléatoire*	3	∞	0.234	13.07	1.50
2887	engin	spatial	*engins spatiaux + engin spatial*	57	∞	0.392	16.94	2.21
4453	étude	poussées	*études poussées*	2	∞	0.236	11.46	0.00
2511	événement	sportif	*événements sportifs*	2	∞	0.346	12.56	0.00
327	fibre	optique	*fibres optiques + fibre optique*	9	∞	0.608	15.80	1.20
3612	focalisation	magnétique	*focalisation magnétique*	2	∞	0.181	12.56	0.96
2909	fonctionnement	anormal	*fonctionnement anormal*	2	∞	0.115	9.39	0.00
3697	génie	civil	*génie civil*	6	∞	0.612	15.36	1.00
1198	gradient	thermique	*gradient thermique + gradients thermiques*	2	∞	-0.177	9.36	1.92
4435	groupe	électrogène	*groupe électrogène + groupes électrogènes*	2	∞	0.120	9.50	0.00
2378	hélium	gazeux	*hélium gazeux*	3	∞	0.711	14.95	0.00
4365	île	salomon	*îles salomon*	4	∞	0.750	15.36	0.00
3926	interconnectivité	total	*interconnectivité totale*	2	∞	-0.203	8.90	3.06
2554	interrogation	préalable	*interrogation préalable*	2	∞	0.118	12.19	1.15
1274	invar	mince	*invar mince*	2	∞	0.463	13.78	0.64
3597	isolation	phonique	*isolation phonique*	2	∞	0.528	13.78	0.00
3320	langue	maternel	*langue maternelle*	2	∞	0.409	13.04	0.00
403	liaison	hétérodyne	*liaison hétérodyne*	2	∞	0.055	7.25	0.00
2280	liste	complet	*liste complète*	2	∞	-0.059	10.84	2.60
770	logique	combinatoire	*logique combinatoire*	2	∞	0.457	13.36	0.00
1620	mélangeur	double		3	∞	0.289	13.36	1.68
1978	microstations	distant	*microstations distantes*	2	∞	-0.178	9.34	1.35
2045	mission	multiple	*missions multiples*	2	∞	-0.249	7.84	1.74
3277	mode	semi-continu		2	∞	0.148	10.11	0.00
2651	module	défectueux	*modules défectueux*	2	∞	0.323	12.36	0.00
1372	mois	quelconque	*mois quelconque*	12	∞	0.548	15.89	1.92
1696	monture	polaire	*monture polaire + montures polaires*	2	∞	0.354	13.36	1.04
1661	navette	spatial	*navette spatiale*	5	∞	-0.088	9.92	2.21
3392	niveau	hiérarchique	*niveaux hiérarchiques + niveau hiérarchique*	4	∞	0.159	10.89	0.00
1351	note	relatif	*notes relatives*	2	∞	-0.202	8.92	**3.43**
4193	observation	général	*observations générales*	15	∞	0.188	13.96	3.58
4204	océan	atlantique	*océan atlantique*	2	∞	0.457	13.36	0.00
4123	océan	indien	*océan indien*	2	∞	0.457	13.36	0.00
1537	octet	indicateur	*octet indicateur*	2	∞	0.354	13.36	0.69
2384	opération	réversible	*opération réversible*	2	∞	0.195	10.90	0.00
2678	oscillation	journalier	*oscillation journalière*	2	∞	0.224	12.78	1.56
3943	pôle	mécanique	*pôle mécanique*	2	∞	-0.091	10.50	2.98
2629	palier	central	*palier central*	2	∞	-0.206	8.84	2.05
2063	paragraphe	suivant	*paragraphes suivants*	4	∞	-0.067	10.47	**3.80**
2571	partage	interrégional	*partage interrégional*	3	∞	0.466	13.73	0.00
2094	partie	gauche	*partie gauche*	2	∞	0.148	10.11	0.00
1375	partie	intégrant	*partie intégrante*	8	∞	0.378	14.11	0.00
3258	personnel	compétent	*personnel compétent*	2	∞	0.210	11.11	0.00
3591	perte	ohmiques	*pertes ohmiques*	2	∞	0.210	11.11	0.00
970	point	nodal	*point nodal*	2	∞	0.222	11.28	0.00
3083	polariseur	quart	*polariseur quart*	2	∞	0.346	12.56	0.00
3480	pondération	psophométrique	*pondération psophométrique*	4	∞	0.566	14.78	0.64
4376	porteur	modulé	*porteuses modulées*	2	∞	0.097	8.90	0.00
346	préjudice	économique	*préjudice économique*	7	∞	0.302	14.12	2.64
2381	proportion	important		2	∞	-0.192	9.10	**3.57**
4219	publication	anticipé	*publication anticipée*	8	∞	0.823	16.36	0.00

Index	N	Adj	R. expression	NC	LOG	FAG	IM3	h2
3990	puissance	rayonné	*puissance rayonnée*	2	∞	0.086	8.54	0.00
1233	référence	bibliographique	*références bibliographiques*	3	∞	0.390	13.21	0.00
738	région	océanique	*régions océaniques*	2	∞	0.236	11.46	0.00
2487	répartiteur	numérique	*répartiteur numérique*	2	∞	-0.281	6.77	3.66
587	réponse	impulsionnelle	*réponse impulsionnelle*	8	∞	0.622	15.56	0.00
284	république	démocratique	*république démocratique*	8	∞	0.432	14.50	0.00
881	république	populaire	*république populaire*	9	∞	0.464	14.84	0.00
4446	réseau	maillés	*réseaux maillés*	5	∞	0.132	10.57	0.00
3820	rôle	prédominant	*rôle prédominant*	2	∞	0.289	12.04	0.00
598	rafraîchissement	conditionnel	*rafraîchissement conditionnel*	7	∞	0.811	16.17	0.00
2067	rdcp	classique	*rdcp classiques*	4	∞	0.074	12.11	2.75
461	séquence	clef	*séquence clef*	3	∞	0.318	12.63	0.00
3	satellite	artificiel	*satellite artificiel + satellites artificiels*	3	∞	0.104	9.39	0.00
1371	satellite	géostationnaires		26	∞	0.387	15.62	0.00
3010	secousse	sismique	*secousses sismiques*	2	∞	0.354	13.36	1.04
294	signal	discret	*signal discret*	3	∞	0.088	8.93	0.00
196	signal	original	*signal original + signaux originaux*	2	∞	0.066	7.76	0.00
3734	sou-systèmes	utilisateurs	*sous-systèmes utilisateurs*	3	∞	0.466	13.73	0.00
2992	soutien	logistique	*soutien logistique*	2	∞	0.646	14.36	0.00
16	srs	communautaire	*srs (communautaire*	2	∞	-0.166	9.53	2.93
3126	stabilisation	triaxiale	*stabilisation triaxiale*	7	∞	0.811	16.17	0.00
4065	stabilité	dimensionnel	*stabilité dimensionnelle*	2	∞	0.528	13.78	0.00
3029	station	brouilleuses	*stations brouilleuses*	4	∞	0.048	7.45	0.00
2201	symétrie	axial	*symétrie axiale*	2	∞	0.181	12.56	0.60
3751	système	sous-régionaux	*systèmes (nationaux et + régionaux ou) sous-régionaux*	2	∞	0.064	7.70	0.00
296	terme	dépendant	*terme dépendant*	5	∞	0.656	15.20	0.00
1508	tirant	parti	*tirant parti*	3	∞	0.577	14.53	0.56
1129	titre	indicatif	*titre indicatif*	3	∞	0.356	12.95	0.00
1145	total	disponible	*total disponible*	6	∞	0.118	12.67	3.11
268	trafic	sporadique		4	∞	0.212	11.72	0.00
3328	transaction	préliminaire	*transaction préliminaire*	2	∞	0.024	11.56	2.04
1684	transition	brutal	*transition brutale + transitions brutales*	2	∞	0.528	13.78	0.00
2051	transposition	régénérateur		2	∞	0.279	13.04	1.05
1707	travail	intérimaire	*travail intérimaire*	3	∞	0.466	13.73	0.00
1526	usage	exclusif	*usage exclusif*	2	∞	0.323	12.36	0.00
32	voisinage	immédiat	*voisinage immédiat*	2	∞	0.224	12.78	1.56
3643	vue	éclatée	*vue éclatée*	2	∞	0.409	13.04	0.00
3575	zone	hydrométéorologiques	*zones hydrométéorologiques*	4	∞	0.234	12.01	0.00
2524	zone	radioclimatiques	*zones radioclimatiques*	2	∞	0.143	10.01	0.00