Study and Implementation of Combined Techniques for Automatic Extraction of Terminology

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

Abstract

This paper presents an original method and its implementation to extract terminology from corpora by combining linguistic filters and statistical methods. Starting from a linguistic study of the terms of telecommunication domain, we designed a number of filters which enable us to obtain a first selection of sequences that may be considered as terms. Various statistical scores are applied to this selection and results are evaluated. This method has been applied to French and to English, but this paper deals only with French.

Introduction

A terminology bank contains the vocabulary of a technical domain: terms, which refer to its concepts. Building a terminological bank requires a lot of time and both linguistic and technical knowledge. The issue, at stake, is the automatic extraction of terminology of a specific domain from a corpus. Current research on extracting terminology uses either linguistic specifications or statistical approaches. Concerning the former, [Bourigault, 1992] has proposed a program which extracts automatically from a corpus sequences of lexical units whose morphosyntax characterizes maximal technical noun phrases. This list of sequences is given to a terminologist to be checked. For the latter, several works ([Lafon, 1984], [Church and Hanks, 1990], [Calzolari and Bindi, 1990], [Smadja and McKeown, 1990]) have shown that statistical scores are useful to extract collocations from corpora. The main problem with one or the other approach is the "noise": indeed, morphosyntactic criteria are not sufficient to isolate terms, and collocations extracted thanks to statistical methods belong to various types of associations: functional, semantical, thematical or uncharacterizable ones.
Our goal is to use statistical scores for extracting technical compounds only and to forget about the other types of collocations. We proceed in two steps: first, apply a linguistic filter which selects candidates from the corpus; then, apply statistical scores to rank these candidates and select the scores which fit our purpose best, in other words scores that concentrate their high values to terms and their low values to co-occurrences which are not terms.

Linguistic Data

In a first part, we therefore study the linguistic specifications on the nature of terms in the technical domain of telecommunications for French. Then, taking into account these linguistics results, we present the method and the program which extracts and counts the candidate terms.

Linguistic specifications

Terms are mainly multi-word units of nominal type that could be characterized by a range of morphological, syntactic or semantic properties. The main property of nominal terms is the morphosyntactic one: its structure belongs to well-known morphosyntactic structures such as N ADJ, N_1 *de* N_2, etc. that have been studied by [Mathieu-Colas, 1988] for French. Some graphic indications (hyphen), morphological indications (restrictions in flexion) and syntactic ones (absence of determiners) could also be good clues that a noun phrase is a term. We have also employed a semantic criteria: the criterion of unique referent. A term refers to an unique and universal concept. However, it is not obvious to apply this criterion to a technical domain where we are not expert. So, we have interpreted the criterion of unique referent by the one of unique translation. A French term is always identically translated, mostly by a compound or a simple noun in English. We have extracted manually terms following these criteria from our bilingual corpus, available in French and English, the *Satellite Communication Handbook (SCH)* containing 200 000 words in each language. Then, we have classified terms following their lengths; the length of a term is defined as the number of *main items* it contains.[1] From this classification, it appears that terms of length 2 are by far the most frequent ones. As statistical methods ask for a good representation in number of the samples, we decided to extract in a first round only terms of length 2 that we will call *base-term* which matched a list of previously determined patterns:

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

N_1 *de* (DET) N_2 *zone de couverture* (*coverage zone*)

N_1 *à* (DET) 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)*), *[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 → N_1 *et* N_2 *de* N_3

   assemblage de paquet + *désassemblage de paquets* → *assemblage et désassemblage de paquets* (*packet assembly/desassembly*)

It is difficult to determine whether a modified or overcomposed base-term is or is not a term. Take for example *bande latérale unique* (single side-band): *bande latérale* (*side-band*) is a base-term of structure N ADJ and *unique* (*single*) a very common post-modifier adjective in French. The fact that *bande latérale unique* is a term is indicated by the presence of the abbreviation *BLU* (*SSB*). As abbreviations are not introduced for all terms, the right attitude is surely to extract first base-terms, i.e. *bande latérale* (*side-band*). Once you have base-terms, you can easily extract from the corpus terms of length greater than 2, at least post-modified base-terms and overcomposed base-terms by juxtaposition. But, even if we have decided to extract only base-terms (length 2), we have to take into account their variations, at least some of them. Variants of base-terms are classified under the following categories:

1. Graphical and orthographic variants

   By graphical variants, we mean either the use or not of capitalized letters (*Service national* or *service national* (*(D/d)omestic service*), or the presence or not of an hyphen inside the N_1 N_2 structure (*mode paquet* or *mode-paquet* (*packet(-)mode*)).

   Orthographic variants concern N_1 PREP N_2 structure. For this structure, the number of N_2 is generally fixed, either singular or plural. However, we

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

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

have encountered some exceptions: *réseau(x) à satellite, réseaux(x) à satellites* (*satellite network(s)*).

2. Morphosyntactic variants
   Morphosyntactic variants refer to the presence or not of an article before the N_2 in the N_1 PREP N_2 structure: *ligne d'abonné, lignes de l'abonné* (*subscriber lines*), to the optional character of the preposition: *tension hélice, tension d'hélice* (*helix voltage*) and to synonymy relation between two base-terms of different structures: for example N ADJ and N_1 *à* N_2: *réseau commuté, réseau à commutation* (*switched network*)
3. Elliptical variants
   A base-term of length 2 could be called up by an elliptic form: for example: *débit* which is used instead of *débit binaire* (*bit rate*).

After this linguistic investigation, we decide to concentrate on terms of length 2 (base-terms) which seem by far the most frequent ones. Moreover, the majority of terms whose length is greater than 2 are built from base-terms. A statistical approach requires a good sampling that base-terms provide. To filter base-terms from the corpus, we use their morphosyntactic structures. For this task, we need a tagged corpus where each item comes with its part-of-speech and its lemma. The part-of-speech is used to filter and the lemma to obtain an optimal sampling. We have use the stochastic tagger and the lemmatizer of the Scientific Center of IBM-France developed by the speech recognition team ([Dérouault, 1985] and [El-Bèze, 1993]).

Linguistic filters

We now face a choice: we can either isolate collocations using statistics and then apply linguistic filters, or apply linguistic filters and then statistics. It is the latter strategy that has been adopted: indeed, the former asks for the use of a window of an arbitrary size; if you take a small window size, you will miss a lot of occurrences, mainly morphosyntactic variants, base-terms modified by an inserted modifier, very frequent in French, and coordinated base-terms; if you take a longer one, you will obtain occurrences that do not refer to the same conceptual entity, a lot of ill-formed sequences which do not characterizes terms, and moreover wrong frequency counts as several short sequences are masked by only one long sequence. Using first linguistic filters based on part-of-speech tags appears as the best solution. Moreover, as patterns that characterizes base-terms can be described by regular expressions, the use of finite automata seems a natural way to extract and count the occurrences of the candidate base-terms.

The frequency counts of the occurrences of the candidate terms are crucial as they are the parameters of the statistical scores. A wrong frequency count implies wrong or not relevant values of statistical scores. The objective is to optimize the count of base-terms occurrences and to minimize the count of incorrect occurrences. Graphical, orthographic and morphosyntactic variants of base-terms (except synonymic variants) are taken into account as well as some syntactic variations that affect the base-terms structure: coordination and insertion of modifiers. Coordination of two base-terms rarely leads to the creation of a new term of length greater than 2, so it is reasonable to think that the sequence *équipements de modulation et de démodulation* (*modulation and demodulation equipments*) is equivalent to the sequence *équipement de modulation et équipement de démodulation* (*modulation equipment and demodulation equipment*). Insertion of modifiers inside a base-term structure does not raise problem, expect when this modifier is an adjective inserted inside a N_1 PREP N_2 structure. Let us examine the sequence *antenne parabolique de réception* (*parabolic receiving antenna*), this sequence could be a term of length 3 (obtained either by over-composition or by modification) or a modified base-term, namely *antenne de réception* modified by the inserted adjective *parabolique*. On one hand, we don't want to extract terms of length greater than 2, but on the other hand, it is not possible to ignore adjective insertion. So, we have chosen to accept insertion of adjective inside N_1 PREP N_2 structure. This choice implies the extraction of terms of length 3 of N_1 ADJ PREP N_2 structure that are considered as terms of length 2. However, such cases are rare and the majority of N_1 ADJ PREP N_2 sequences refer to a N_1 PREP N_2 base-term modified by an adjective.

Each occurrence of a base-terms is counted equally; we consider that there is equiprobability of the term appearance in the corpus. The occurrences of morphological sequences which characterize base-terms are classified under pairs: a pair is composed of two main items in a fixed order and collects all the sequences where the two lemmas of the pair appear in one of the allowed morphosyntactic patterns; for example, the sequences: *ligne d'abonné, lignes de l'abonné* (*subscriber lines*), *ligne numérique d'abonné* (*digital subscriber line*) are each one occurrence of the pair (ligne, abonné). If we have the coordinated sequence *lignes et services d'abonné* (*subscriber lines and services*), we count one occurrence for the pair (ligne, abonné) and one occurrence for the pair (service, abonné). Our program scans the corpus and counts and extracts collocations whose syntax characterizes base-terms. Under each pair, we find all the different occurrences found with their frequencies and their location in the corpus (file, sentence, item). This program runs fast: for example, it took 2 minutes to extract 8 000 pairs from our corpus *SCH* (200 000 words) for the structure N_1 *de* (DET) N_2 on a Sparc station ELC (SS1) under Sun-Os Release 4.1.3.

Now that we have obtained a set of pairs, each pair representing a candidate term, we apply statistical scores in order to distinguish terms from non-terms among the candidates.

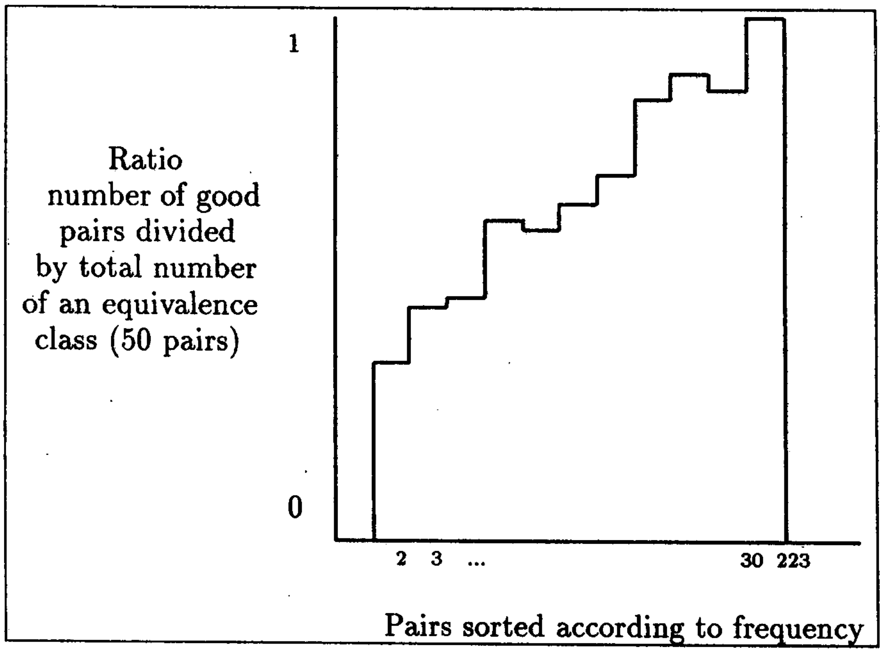

Figure 1: Frequency histogram

Lexical Statistics

The problem to solve now is to discover which statistical score is the best to isolate terms among our list of candidates. So, we compute several measures: frequencies, association criteria, Shannon diversity and distance scores. All these measures could not be used for the same purpose: frequencies are the parameters of the association criteria, association criteria propose a conceptual sort of the couples, and Shannon diversity and distance measures are not discriminatory scores but provide other types of informations.

Frequencies and 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 statistical literature proposes many scores which can be used to test the strength of the bond between the two variables of a contingency table. Some are well-known such as the association ratio, close to the concept of mutual information, introduced by [Church and Hanks, 1990]:

$$IM = \log_2 \frac{a}{(a+b)(a+c)} \tag{1}$$

the Φ^2 coefficient introduced by [Gale and Church, 1991]:

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

or the Loglike coefficient introduced by [Dunning, 1993]:

$$\begin{aligned} Loglike = {} & a\log a + b\log b + c\log c + d\log d \\ & -(a+b)\log(a+b) - (a+c)\log(a+c) \\ & -(b+d)\log(b+d) - (c+d)\log(c+d) \\ & +(a+b+c+d)\log(a+b+c+d) \end{aligned} \tag{3}$$

A property of these scores is that their values increase with the strength of the bond of the lemmas. We have tried out several scores (more than ten) including IM, Φ^2 and Loglike and we have sorted the pairs following the score value. Each score proposes a conceptual sort of the pairs. This sort, however, could put at the top of the list compounds that belong to general language rather than to the telecommunication domain. As we want to obtain a list of telecommunication terms, it is

Pairs of N_1 (PREP (DET)) N_2 structure	The most frequent pair sequence	Logl	Nbc	IM
(largeur, bande)	***largeur de bande*** (197) (***bandwidth***)	**1328**	223	5.74
(température, bruit)	***température de bruit*** (110) (***noise temperature***)	**777**	126	6.18
(bande, base)	***bande de base*** (142) (***baseband***)	**745**	145	5.52
(amplificateur, puissance)	***amplificateur(s) de puissance*** (137) (***power amplifier***)	**728**	137	5.66
(temps, propagation)	***temps de propagation*** (93) (***propagation delay***)	**612**	94	6.69
(règlement, radiocommunication)	***règlement des radiocommunications*** (60) (***radio regulation***)	**521**	60	8.14
(produit, intermodulation)	***produit(s) d'intermodulation*** (61) (***intermodulation product***)	**458**	61	7.45
(taux, erreur)	***taux d'erreur*** (70) (***error ratio***)	**420**	70	6.35
(mise, œuvre)	***mise en œuvre*** (47) (***implementation***)	**355**	47	7.49
(télécommunication, satellite)	***télécommunication(s) par satellite*** (88) (***satellite communication(s)***)	**353**	99	4.09
(bilan, liaison)	***bilan(s) de liaison*** (37) (***link budget***)	**344**	55	6.42

Figure 2: Topmost pairs

essential to evaluate the correlation between the score values and the pairs and to find out which scores are the best to extract terminology. Therefore, we compare the values obtained for each score to a reference list of the domain. We have used the terminology data bank of the EEC, telecommunication section, which has been elaborated by experts. This evaluation has been done for 2 200 French pairs[3]of N_1 *de* (DET) N_2 structure extracted from our corpus *SCH* (200 000 words). Each score provides as a result a list where the candidates are sorted following the score value. We have defined equivalence classes which generally collect 50 successive pairs of the list. The results of a score are represented graphically thanks to an histogram in which the x-axis represents the pairs sorted according to the score value, and y-axis the ratio of the number of pairs belonging to the reference list divided by the number of pairs per equivalence class, i.e. generally 50 pairs. If all the pairs of an equivalence 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 its high values (resp. low) to good (resp. bad) pairs, i.e. candidates which belong (resp. which 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). We are not going to present here all the histograms obtained (see [Daille, 1994]). All of them show a general growing trend that confirm that the score values increase with the strength of the bond of the lemma. However, the growth is more or less clear, with more or less sharp variations. The most beautiful histogram is the simple frequency of the pair (see Figure 1). This histogram shows that more frequent the pair is, the more likely the pair is a term. Frequency is the most significant score to detect terms of a technical domain. This results contradicts numerous results of lexical resources, which claim that association criteria are more significant than frequency: for example, all the most frequent pairs whose terminological status is undoubted share low values of association ratio (formula 1) as for example *réseau à satellites* (*satellite network*) IM=2.57, *liaison par satellite* (*satellite link*) IM=2.72, *circuit téléphonique* (*telephone circuit*) IM=3.32, *station spatiale* (*space station*) IM=1.17 etc. The remaining problem with the sort proposed by frequency is that it integrates very quickly bad candidates, i.e. pairs which are not terms. So, we have preferred to elect the Loglike coefficient (formula 3) the best score. Indeed, Loglike coefficient which is a real statistical test, takes into account the pair frequency but accepts very little noise for high values. To give an element of comparison, the first bad candidate with frequency for the general pattern N_1 (PREP (DET)) N_2 is the pair (cas, transmission) which appears in 56th place; this pair, which is also the first bad candidate with Loglike, appears in 176th place. We give in figure 2 the topmost 11 french pairs sorted by the Loglike coefficient (Logl) (Nbc is the number of the pair occurrences and IM the value of

[3] Only pairs which appear at least twice in the corpus have been retained.

association ratio).

Diversity

Diversity has been introduced by [Shannon, 1948] and characterizes the marginal distribution of the lemma of a pair through the range of pairs. Its computation uses a contingency table of length n: we give below as an example the contingency table which is associated to the pairs of N ADJ structure:

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

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:

$$\begin{aligned} H_i &= nb_{i.} \log n_{i.} - \sum_{j=1}^{s} nb_{ij} \log nb_{ij} \qquad (4) \\ H_j &= nb_{.j} \log n_{.j} - \sum_{i=1}^{s} nb_{ij} \log nb_{ij} \end{aligned}$$

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

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

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:

$$\begin{aligned} h_i &= \frac{Hi}{n_{ij}} \qquad (5) \\ h_j &= \frac{Hj}{n_{ij}} \end{aligned}$$

The normalized diversities h_1 and h_2 are defined from H_1 and H_2.

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. High values of h_1 applied to the pairs of N ADJ structure characterizes nouns that could be seen as key-words of the domain: *réseau* (*network*), *signal*, *antenne* (*antenna*), *satellite*. Conversely, high values of h_2 applied to the pairs of N ADJ structure characterizes adjectives which do not take part to base-MWUs as *nécessaire* (*necessary*), *suivant* (*following*), *important*, *différent* (*various*), *tel* (*such*), etc. The pairs with a zero diversity on one of their lemma receive high values of association ratio and other association criteria and a non-definite value of Loglike coefficient. However, the diversity is more precise because it indicates if the two lemmas appear only together as for (**océan, indien**) (*indian ocean*) (H_1=h_1=H_2=h_2=0), or if not, which of the two lemmas appear only with the other, as for (**réseau, maillé**) (*mesh network*) (H_2=h_2=0), where the adjective *maillé* appears only with *réseau* or for (**codeur, idéal**) (*ideal coder*) (H_1=h_1=0) where the noun *codeur* appears only with the adjective *idéal*. Other examples are: (**île, salomon**) (*solomon island*), (**hélium, gazeux**) (*helium gas*), (**suppresseur, écho**) (*echo suppressor*). These pairs collects many frozen compounds and collocations of the current language. In future work, we will investigate how to incorporate the nice results provided by diversity into an automatic extraction algorithm.

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:

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

$$\sigma(X) = \sqrt{V(X)}$$

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. 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.

1. **Pairs with no distance variation** $V(X) = 0$

   (a) N_1 N_2 : *Dist* = 2 *MDist* = 2

   - *liaison sémaphore, liaisons sémaphores* (*common signalling link(s)*)
   - *canal support, canaux support, canaux supports* (*bearer channel*)

(b) N_1 PREP N_2 : $Dist = 3$ $MDist = 2$

- *accusé(s) de réception* (*acknowledgement of receipt*)
- *refroidissement à air, refroidissement par air* (*cooling by air*)

(c) N_1 PREP DET N_2 : $Dist = 4$ $MDist = 2$

- *sensibilité au bruit* (*susceptibility to noise*)
- *reconnaissance des signaux* (*signal recognition*)

(d) N_1 ADJ PREP 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 $V(X) \neq 0$**

- (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

Conclusion

We presented a combining approach for automatic term extraction. Starting from a first selection of lemma pairs representing candidate terms from a morphosyntactic point of view, we have applied and evaluated several statistical scores. Results were surprising: most association criteria (for example, mutual association) didn't give good results contrary to frequency. This bad behavior of the association criteria could be explained by the introduction of linguistic filters. We can notice anyway that frequency characterizes undoubtedly terms, contrary to association criteria which select in their high values frozen compounds belonging to general language. However, we preferred to elect the Loglike criterion rather than frequency as the best score. This latter takes into account frequency of the pairs but provide a conceptual sort of high accuracy. Our system which uses finite automata allows to increase the results of the extraction of lexical resources and to demonstrate the efficiency to incorporate linguistics in a statistic system. This method has been extended to bilingual terminology extraction using aligned corpora ([Daille *et al.*, 1994]).

Acknowledgments

I would like to thank the IBM-France team, and in particular Éric Gaussier and Jean-Marc Langé, for the tagged and lemmatized version of the French corpus and for their evaluation of statistics; Owen Rambow for reviewing. This research was supported by the European Commission and IBM-France, through the *ET-10/63* project.

References

[Bourigault, 1992] Didier Bourigault. 1992. Surface grammatical analysis for the extraction of terminological noun phrases. In *Proceedings of the Fourteenth International Conference on Computational Linguistics (COLING-92)*, Nantes, France.

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

[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.

[Dérouault, 1985] Anne-Marie Dérouault. 1985. *Modélisation d'une langue naturelle pour la désambiguation des chaînes phonétiques*. PhD thesis, University Paris VII, France.

[Dunning, 1993] Ted Dunning. 1993. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, vol. 19, n° 1.

[El-Bèze, 1993] Marc El-Bèze. 1993. *Les Modèles de Langage Probabilistes : Quelques Domaines d'Applications*. Habilitation à diriger des recherches (Thesis required in France to be a professor), University Paris-Nord, France.

[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.

[Gale and Church, 1991] William A.Gale and Kenneth W.Church. 1991. 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.

[Mathieu-Colas, 1988] Michel Mathieu-Colas. 1988. Typologie des noms composés, Technical report n° 7, Paris, *Programme de Recherches Coordonnées "Informatique et Linguistique"*, University Paris 13.

[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.