Contrasting French nominal terms to common language NPs – towards a rule-based term extractor

Ágoston Nagy

The aim of the paper is (1) to point out the specificities of French nominal terms with respect to prepositional complements and adjectival adjuncts; and (2) to contrast the internal structure of nominal terms with that of common language NPs. This analysis is used to elaborate the rule-based module of an automatic term extractor, the main aim of which is to find nominal terms in a specialised text and to filter out common language nominal expressions.
This rule-based module is complemented with a rule-based filter. The corpora used for the analysis consist of the description parts of patents written in French.
As hypothesised, the results showed that already with rule based methods, high accuracy can be achieved. Without the rule-based filtering, the program produced high recall (0,82) and low precision (0,53). The filtering resulted in the increase in both the precision (0,66) and the recall (0,83).

Keywords: *adjectival adjunct, nominal phrase, nominal term, prepositional complement, term extraction*

1 Introduction

According to the traditional approach (e.g. Wüster 1976), terms are lexical units having the following characteristics: they are related to a domain (e.g. informatics, physics), they are connected to one and only one concept that they denote. According to Justeson and Katz (1995), in computational terminology, nominal terms are in the centre of interest since these are the terms having the most complex syntactic structure ranging from simple nouns to extremely long nominal compositions (e.g. in French *donnée* 'data' and *système de gestion de base de données* 'database management system').

The aim of this article is to give an overview on the differences between noun phrases (NPs) and nominal terms, especially with respect to prepositional complements (including the preposition+noun sequences of nominal compounds) and adjectival adjuncts. This research is done for the purpose of creating a term extractor, which is a program used to automatically extract nominal terms from written, French language patent texts. As the main aim of a nominal term extractor is to extract nominal terms, and therefore exclude common language noun phrases, the internal structure of nominal terms will be contrasted to that of ordinary NPs. This contrastive analysis will be based on French grammar books and articles about the structure of NPs, like Riegel *et al* (2009) and Anscombre (1991), respectively.

Throughout the article, the notion of *term* is used for the elements to be analysed in this paper and to be extracted by the application even though the domain-relatedness of these elements is not taken into consideration when their internal structure is presented in this paper and implemented into the application. It is hypothesised that the choice of the corpus determines their domain-relatedness, and as the corpus itself highly represent specialised discourse, it does not contain terms of other domains. The analysed

Proceedings of the First Central European Conference in Linguistics for postgraduate Students (2012), 191–210.
ISBN 978-963-308-054-2

elements cannot be considered as *concepts* because concepts are the abstract mental representations of terms.

Since this automatic term extractor works on predefined syntactic patterns in order to extract nominal candidate terms, the different syntactic structures of terms have to be defined and the differences between common language NPs and terms have to be shown, and both of them as precisely as possible. In this article it is the traditional definition of terms (Würster 1976, Cabré 1999) that is used that claims that a term denotes only one concept, and as such, it has to form one lexical unit that resembles to a simple noun or a nominal compound. On the contrary, an NP is a major category, a syntactic unit having an obligatory head to which different complements, adjuncts or determiners are attached.

Out of the three examples in (1), it is only (1a) that can be a term (e.g. in the domain of public administration) and an NP in the same time, the others (1b,c) only being NPs.

(1) a. *Un hôtel de ville*
a hotel of town
'townhall'
b. *un hôtel de la ville*
'a hotel of the town'
c. *le rat noir qui fait la sieste*
'the black rat having a rest'

A term can dispose of other supplementary elements, like adjectives, but only some kinds of adjectives can be part of terms: only those that designate a subclass of the nominal head, like *binaire* 'binary' in (2a). As (2b) does not represent a subclass of files, the adjective *gros* 'big, fat' does not form a part of the term.

(2) a. *un fichier binaire*
'a binary file'
b. *un gros fichier*
'a big file'

In order that terms and NPs could be differentiated, the basic structure of French NPs will be presented especially with respect to prepositional phrases (PP) and adjectival phrases (AP) that can be adjoined to the nominal head, because these are the two where the difference between NPs and nominal terms are more pertinent.

The other aim of the article is to present the results of the term extractor I elaborated on the basis of the differences found between PPs and APs that can be adjoined to the nominal head. The term extractor also uses different filtering techniques, one rule-based and three statistical ones that can filter out elements that are not, or may not be part of terms (see Section 2).

In the second section, the different methods of term extraction will be described as well as the method of my own term extractor. In Section 3, the corpus will be presented: the latter contains patent descriptions of two scientific domains, namely informatics and human necessities. This is followed by the presentation of the two basic NP constituents, APs and PPs (Section 4), which precedes the analysis of these constituents in nominal terms (Section 5). In Section 6, the results of the automatic term extractor, the used syntactic patterns of which are based on the previous sections, will be

presented. The main aim of the next and last section is to present which are the possible sources of error of this automatic term extractor, and to find out whether these errors are due to the syntactic patterns or not.

2 Term extraction – methods, hypotheses

Terminology extraction, just like any other domain of computational linguistics, can be realised by rule-based and statistical methods, but this does not mean that these applications only use one of the two: most of them rely on both methods (Maynard and Ananiadou 2001). According to Cabré et al. (2001), it is not recommended to use only one of them, because rule-based methods result in too much noise (i.e. the number of extracted terms is bigger than that of real terms), whereas statistical methods provoke too much silence (the list of extracted terms does not contain many of the real terms). In term extraction, rule-based methods mean that terms can be extracted on the basis of their inner syntactic structure, for example if a noun is followed by another one, the whole can be marked as a term. Statistical methods mean that we look for example for sequences that occur more times in a specific corpora than in general language: these can then probably be marked as a term.[1]

According to Cabré et al. (2001), the best term extraction tools extract first candidate terms by means of statistical methods, and this list is then filtered with linguistic filters. However, in my term extraction tool, I chose the inverse direction: stopwords were firstly filtered out from the text by rule-based filters (later referred to as RBF), and then terms were extracted on the basis of their internal syntactic structure by rule-based extraction (later referred to as RBE). And as an experiment, this list was filtered with statistical methods (later referred to as SF), too.

The rule-based filtering consists of deleting nominal and adjectival stopwords from the text that cannot be part of terms. These elements are mainly connectives, that is their function is limited to provide the cohesion of a text, like *en effet* 'in fact' or *par exemple* 'for example'. These have to be filtered out because these expressions containing at least one noun cannot be or cannot be part of terms. The stopword list also comprises adjectives that has the same function, i.e. providing text cohesion: these are for example *suivant* 'following' or *précédent* 'previous'. That is, if the text contains the expression [*les*] *acides gras suivants* '[the] following fatty acids', it will be reduced to *acides gras* 'fatty acids'. Rule-based filtering takes place before rule-based extraction, therefore these stopwords are not deleted from the candidate terms themselves.

The rule-based extraction module uses a finite state automaton to recognise nominal terms. This automaton was manually created on the basis of grammar rules describing the characteristics of nominal compounds and on the basis of the findings of this article.

As an experiment, a combination of statistical methods will also be used to furthermore filter out the candidate term list. These statistical methods include firstly the

[1] The extraction of units having a high frequency as compared to other elements in a text suggest that it is collocations and not terms that are in the centre of the analysis. However, collocations represent a much broader category than terms since (1) one-word terms cannot be considered as collocations and (2) the chosen corpus contains many collocations that cannot be terms, e.g. *La présente invention concerne …* 'The present invention concerns ...' where *présente* 'present' is not part of a term, and thus have to be filtered out.

weirdness value (Ahmad et al. 1999) of which the main aim is to compare the frequency of candidate terms in the specialised corpora to their frequency in a common language corpora. In fact, weirdness is calculated in the term extractor as the proportion of the relative frequency of the candidate term in specialised discourse and its relative frequency in common language. The second one, the weight value (Frantzi & Ananiadou 1997), consists of assigning every candidate term a probabilistic value based on its textual environment (e.g. expressions preceded by *est appelé* 'is called' are more likely to become terms). It is a statistical algorithm which assigns to every word in the corpus a probabilistic value, which is high if it mostly follows or precedes terms and low if it is rarely in the environment of terms. The third one, the C-value (Frantzi & Ananiadou 1997), is used to measure the inner cohesion of the constitutive elements of a complex nominal term: for example, if the constitutive elements tend to be used separately more often, then it is these separate parts that are more probable to become terms, thus they get a higher value. This can be computed by calculating the frequency of the candidate terms and their parts apart.[2]

This article first reveals how efficient our contrastive analysis is since the patterns gained in this way will be used in a term extractor based on nominal term patterns: my hypothesis is that rule-based approach to term extraction from French patent corpora can already be efficient without statistical methods since (1) in French, terms tend to have internal structures that are not typical of common language nouns or nominal compositions and (2) patents represent a discourse type that corresponds to nearly all prerequisites of a scientific text (e.g. impersonal style, excessive usage and repetition of terms).

3 Corpus[3]

French language patents were chosen as the corpus of the analysis, because patents are written in a way to comply with the prerequisites of a specialised text, and terms can only be extracted from specialised corpora. A patent is divided into many units, like bibliographical data, summary, description and claims. From among these parts, our analysis is restricted to the description part of patents because (1) the description part is the most detailed and the longest part of a patent enumerating the advantages of the new invention and (2) as the description has to be as precise as possible, terms are frequently repeated in it as such without any modification. This leads us to the presupposition that even statistical methods can work well on these texts.

In our analysis, focus is given on patents of two different domains: one is the G06F patent class dealing with informatics and the other is the A23L class which represents the Human necessities domain. From these two areas, ten descriptions were chosen as samples on which the application was executed. In order to measure the effectiveness of the rule-based extraction, as well as of the rule-based and statistical methods, terms have manually been annotated, that is they have been marked as terms in these descriptions. Consequently, the term extractor can compare the list provided by

[2] These metrics are not presented in detail, because their application is in an experimental stage, and they do not make real part of the present article.

[3] Although the best placement for Section 3 would be after Section 5, the former was chosen to be the third section as the latter often make reference to the corpus described in Section 3.

itself and that of the previously annotated text. In the G06F corpus the manual annotation found 1752 terms, and in the A23L corpus this number was 2086[4].

In order to demonstrate the different error sources in Section 7, one description was chosen from each of the two corpora. Whenever a specific counterexample is found during the analysis or if the error rate of a specific problematic case is given, it will be based on these two texts, named together *example corpus*. The title of these texts are the following:

A23L: *Use of saffron and/or safranal and/or crocin and/or picrocrocin and/or derivatives thereof as a sateity agent for treatment of obesity*[5]

G06F: *Data exchange between an electronic payment terminal and a maintenance tool through a USB link*[6]

4 The distribution of adjectives and prepositional complements in NPs

The aim of this section is to present in more details the specificities of general NPs, laying stress on its prepositional complements and adjectival adjuncts.

4.1 The distribution of adjectives in NPs

As Cinque (1998) states, French, like most of Romance languages, is an ANA language meaning that adjectives can either precede or follow the nominal head in a NP (e.g. (3a)). On the contrary, Germanic languages, like English and German, are AN languages, that is adjectives can only precede the nominal head (6b,c):

(3) a. *la jolie chambre bleue*
the nice room blue
'the nice blue room'
b. *the nice blue room*
c. ** the nice room blue*

This statement implies that all adjectives could be placed either before or after the nominal head in French, which is not true in all cases. The default position of adjectives is the postnominal position since many adjectives (like *bleue* 'blue' in (3a)) cannot even precede the noun. According to Riegel et al. (2009), on average one adjective out of three is placed before the noun but there can be enormous differences between different types of discourse (in literary language one adjective out of two is before the noun but this proportion is one out of ten in scientific language).

[4] Although the manual annotation was carried out by one person – which is normally not recommended in computational linguistics – I did rely on terminological resources (e.g. *Le grand dictionnaire terminologique*: http://www.granddictionnaire.com/btml/fra/r_motclef/index800_1.asp) when annotating the terms in the texts.

[5]Source: http://www.wipo.int/patentscope/search/en/detail.jsf?docId=WO2007125243&recNum=1&maxRec=&office=&prevFilter=&sortOption=&queryString=&tab=PCTDescription

[6]Source: http://www.wipo.int/patentscope/search/en/detail.jsf?docId=WO2009053626&recNum=1&maxRec=&office=&prevFilter=&sortOption=&queryString=&tab=PCTDescription

From among the different adjective types, it is non-classifying adjectives[7] that can either precede or follow the noun. The default place of these adjectives is the postnominal position but these can precede the noun in case they are accentuated or if they are, in other words, focalised. (Laenzlinger 2003)

(4) a. *un roman ennuyeux*
a novel boring
'a boring novel'
b. *un ennuyeux roman*
'a boring novel'

In certain cases, there is a certain semantic difference between the prenominal and postnominal adjective. According to Bouchard (1998), adjectives following the nominal head seem to modify the semantic components of the noun as a whole whereas the same adjective, used as prenominal, tend to modify the inner semantic components of the noun. (5) and (6) show typical cases where the prenominal adjective does not mean the same as its postnominal version:

(5) a. *mon fauteuil ancien*
my armchair old
'my old armchair'
b. *mon ancien fauteuil*
'my old armchair'

(6) a. *un parent seul*
'a lonely parent'
b. *un seul parent*
'only one parent'

These examples clearly show the difference in the semantic interpretation of the pre- and post-nominal variant of the adjectives. For instance, (5a) refers to an 'armchair produced long time ago' whereas (5b) refers to an 'armchair that was not necessarily produced long time ago but which is mine for a long time'.

Some adjectives tend to precede always the noun: these are normally "short", mono- or bi-syllabic adjectives. In this case, grammars also refer to phono-rhythmical and usage factors: these adjectives tend to be frequently used in everyday communication (Laenzlinger 2003). It can easily be understood if one has a look at (7):

(7) a. *une petite chose*
'a little thing'
b. *une belle chanson*
'a beautiful song'
c. *une petite belle tour*
'a little nice tower'

[7] Non-classifying adjectives: adjectives that designate subjective properties (e.g. *nice*) and can be modified by adverbs of degree (*very nice*).

Classifying adjectives: adjectives that designate objective properties (e.g. *black*) and cannot be modified by adverbs of degree (**very black*).

On the basis of (7c), it can be concluded that more than one adjective can be placed before the noun at the same time.

However, if these adjectives are followed by a complement, the AP must be post-nominal. In other words, if the AP has a PP (or any other) complement, it must obligatorily follow the noun.

(8) a. *une longue rivière*
'a long river'
b. *une rivière* $_{AP}$*[longue de 1200 mètres]*
'a 1200 meter long river'
c. *une rivière* $_{AP}$*[moins longue que le Nil]*
'a river shorter than the Nile'
d. **une* $_{AP}$*[longue de 1200 mètres] rivière*
a long of 1200 meters river

A pre-nominal adjective also becomes post-nominal if it is modified by an adverb (10). In fact, this rule does not apply if the adverb is short and frequently used, as *tout* 'completely', *très* 'very' or *trop* 'too': in these cases, the distribution of the AP within the NP is facultative (9b).

(9) a. *une courte enfance*
'a short childhood'
b. *une très courte enfance / une enfance très courte*
'a very short childhood'

(10) a. *une enfance extrêmement courte*
'an extremely short childhood'
b. **une* $_{AP}$*[extrêmement courte] enfance*

And finally, there are some adjectives that can only be post-nominal, these are the so-called intersective predicative adjectives (Bouchard 1998) that normally denote concrete properties, such as origin, shape, colour and the fact to belong to a community. In addition, derived adjectives (such as past or present participles used as adjectives) can only be post-nominal.

(11) a. *le bureau oval*
'the oval office'
b. *un parapluie chinois*
'a Chinese umbrella'
c. *un solvant chimique*
'a chemical solvent'

(12) a. *un loyer modéré*
'a low rent'
b. *un tapis roulant*
'a conveyor belt'

Derived *intensional* adjectives constitute an apparent counter-example, as *intensional* adjectives can only be pre-nominal in French, even if they are derived from a participle, as showed in (13):

(13) a. *un prétendu chef d'Etat*
'a pretended head of state'
b. *un soi-disant dentiste*
'a self-styled dentist'

Another aspect that absolutely has to be taken into consideration is the combination of adjectives and prepositional phrases within nominal expressions. Prepositional phrases do not constitute any problems in this aspect since they obligatorily have to follow the head noun. What is a question that has to be answered is that a post-nominal adjective follows or precedes the prepositional complement. According to Laenzlinger (2003), an adjective can intervene between the head noun and the prepositional phrase but it can follow the complement as well, so its position is facultative, as illustrated by (14):

(14) a. *un minister de la Justice blanc*
a minister of the justice white
'white attorney general'
b. *un ministre blanc de la Justice*

However, if the noun and the prepositional complement form together a lexically fixed entity, adjectives tend not to intervene between them, as shown in (15), where the lexical entity that sticks together is *lunettes de soleil*, literally 'glasses of sun', that is 'sunglasses'.

(15) a. *les lunettes de soleil nouvelles*
the glasses of sun new
'the new sunglasses'
b. *??des lunettes nouvelles de soleil*

Abeillé & Godard (1999) present another approach to the relative position of nouns and adjectives within French nominal expressions. They propose the term *relative weight* in order to give a constraint on the relative ordering of adjectives with respect to nouns. In their terms, the distribution of adjectives within NPs are constrained by their weight. There are two types of weight: "lite" and "non-lite", and this distinction differences lexemes and phrases from each other. Whether an adjective is light or not are either determined by the lexicon (certain adjectives are light, others non-light, and for many adjectives this feature is underspecified), or by the rules describing the syntactic structure of phrases (most of the phrases are non-light, the others light). For example, a rule states that light adjectives must be pre-nominal, non-light ones must be post-nominal.

(16) a. *une* $_{light\ A}$*[belle] dame*
'a beautiful lady'
b. *une femme* $_{non\text{-}light\ A}$*[russe]*
'a Russian woman'

However, light adjectives cannot be adjoined to non-light nouns (such as coordinated nouns): in these cases, they can only be heavy (hence it can only follow the noun), as exemplified in (17):

(17) *non-light N[des hommes et des enfants] non-light A[jolis]*
'nice men and children'

The feature of relative weight of coordinated adjectives is another issue that has to be discussed. The weight of two light coordinated adjectives become underspecified, that is the new AP can either follow or precede the noun (as exemplified in (18)), and two coordinated adjectives for which this feature is lexically not defined become heavy (as exemplified in (19)):

(18) a. *une A non.det.[light-A[jolie] et light-A[belle]] chambre*
'a nice and beautiful room'
b. *une chambre A non.det.[light-A[jolie] et light-A[belle]]*

(19) a. *??une non-light A[A non.det.[excellente] et A non.det.[joyeuse]] femme*
'an excellent and cheerful woman'
b. *une femme non-light A[A non.det.[excellente] et A non.det.[joyeuse]]*

4.2 Prepositional complements in NPs

It is not easy to define the notion of PP in French because it is not always evident whether a PP introduces a new entity inside the NP (20a) or it is associated to the nominal head with which it forms a complex noun (20b)[8]. In the case of nominal compounds (20b), the preposition is in general followed by a noun without a determiner because the presence of a determiner would imply a complex NP where the NP preceded by a preposition could be considered as an embedded NP having a separate reference (20a). However, PPs in nominal compounds are not referential. Hence, determiners have a crucial role in differentiating between *N Prep N* type nominal compounds and NPs having a PP.

(20) a. *le verre de la voisine*
'the glass of the neighbour'
b. *le verre de lait*
the glass of milk
'the milk glass'

Riegel et al. (2009) classes as PP all phrases that is made up of a preposition followed by an entire nominal group (e.g. *le chien PP[de NP[la voisine]] the dog of the neighbour*). In the meanwhile, they also mention nominal compounds like *canne à pêche* 'fishing rod' where *pêche* 'fishing' does not constitute an NP alone. However, Bosredon and Tamba (1991) differentiates the two different prepositional structures: they think that nominal compounds are simple nouns from a semantic point of view but they constitute a NP from a formal point of view. In this way, they distinguish the PPs from the

[8] And therefore it can become a term, e.g. (20b) is a term in the domain of glass fabrication.

preposition+noun sequences that are attached to a noun and they call them constituents and formants, respectively.

In this article, prepositional formants and constituents are both considered as PPs because the boundary between formants and constituents are not so clear from a formal point of view. Firstly, there are prepositions that are not followed by a determiner in general (for example *par* ou *en*) and they do not form a compound noun with the preceding noun (e.g. (21a)). Secondly, there are nominal compounds that contain a PP with a determiner (e.g. (22a)).

(21) a. *voyage en Italie*
'a trip to Italy'
b. **voyage en l'Italie*

(22) a. *maladie de la peau*
'skin illness'
b. *?maladie de peau*

In the remaining part of this section, it is nominal compounds that will be treated in more details because they are the ones that are more likely to become terms.

In French, nominal compounds are created by nouns (23a-d) or infinitives (23e) attached to the nominal head by means of the preposition *de* (23a,b) but in some cases, they can be linked together by other prepositions (like *en* in (23c) or *à* in (23d)). Nominal compounds are written in general without a hyphen, with the exception of some cases, like (23c). (Riegel et al. 2009)

(23) a. *lunettes de soleil*
'sunglasses'
b. *professeur de hongrois*
'teacher of Hungarian'
c. *arc-en-ciel*
'rainbow'
d. *verre à eau*
'water glass'
e. *machine à laver*
'washing machine'

The presence of the hyphen is not only a question of spelling. The automatic term extractor relies on automatic annotations, and these programs (including the one I use for this analysis) does not treat hyphenated elements as different words but as one word the part of speech tag of which is a noun. Hence, nominal compounds like (23c) are recognised by the same syntactic pattern as the one used to recognise terms that are made up of only one noun, like *réseau* 'network' (rules can be found in Section 6.1).

In French, there are also nominal compounds without preposition that can be written with (24a,b) or without (24c) a hyphen.

(24) a. *le gratte-ciel*
'skyscraper'
b. *le chou-fleur*
'cauliflower'

c. *la pause café*
'coffee break'

5 PPs and APs in nominal terms

The aim of Section 5 is to present the possible nominal term structures with respect to adjectival adjuncts and prepositional complements.

5.1 APs in nominal terms

The place of adjectives is a crucial point when it has to be decided whether a specific adjective can appear in a term or not. As it was already mentioned, the default place for adjectives in an NP is the postnominal position but certain adjectives can appear in a prenominal position as well, for example in case of emotional stress. This emotional stress does not play any role in specialised languages since the latter require strong objectivity: it uses only classifying and relational adjectives and thus does not use this affective accentuation. This is exemplified in (25) where the relational adjective cannot be placed before the noun for whatever reason it would be placed before:

(25) a. *un réseau filaire*
a network wired
'a wired network'
b. **un filaire réseau*

Frequently used monosyllabic adjectives have little chance of appearing in a term because they rather designate accidental characteristics of terms. It is the case of the *intensional* adjectives that are derived from a verb (e. g. *prétendu* 'pretended') that can also precede a term but are never part of:

(26) a. *un grand réseau filaire*
a big network wired
'a big wired network'
b. *un prétendu réseau filaire*
a pretended network wired
'a pretended wired network'

In Nagy (2009), it was stated that there was no term that would start with an adjective placed before the noun in an IT corpus. Hence, this possibility will be excluded even if in other terms in other patent domains, there can be some adjectives that is placed before the verb:

(27) a. *petite aiguille*
little needle
'hour hand'
b. *premier ministre*
first minister
'prime minister'

In (27b) the ordinal adjective precedes the noun, like most of the ordinal adjectives, but this type of adjective has mostly an anaphoric role or of text organising, that is it generally refers to a specific occurrence of an already mentioned noun, thus it is usually not a part of the term.

Besides our study, no data on the proportion of terms beginning with an adjective is known that is why, on the basis of the above mentioned study, this possibility will not be taken into consideration. In the two pattern descriptions, there was only one case where an adjective in an embedded PP preceded the nominal head in a term (28):

(28) *acide gras à longue chaîne*
acid fatty with long chain
'long chain fatty acid'

5.2 PPs in nominal terms

From the different complements or adjuncts that an NP can have, it is only the PP or the AP that can appear inside a nominal term. Clauses with a finite verb cannot be part of a term since they generally introduce a new entity in the discourse. For example, in (29), a new concept appears in the relative clause, namely *utilisateur* 'user'.

(29) a. *le site web que l'utilisateur a visité*
'the web site that the user visited'

From the different complements, terms can only have a prepositional complement. As it was already mentioned in the previous sections, lexicalised prepositional nominal compounds in French generally do not contain internal determiners in the PP complement, and the preposition is followed by a noun (the term in (30a)) or an infinitive (30b)).

(30) a. *huile de tournesol*
'sunflower oil'
b. *machine à laver*
'washing machine'

These composed elements have to be considered as a lexical unit because *huile* 'oil' and *tournesol* 'sunflower' are lexical units that can appear as autonomous units but as soon as they are joined together with the preposition *de*, they designate a third different concept. However, if the preposition is followed by a determiner, the situation becomes more complex. By default, a PP complement having a determiner cannot be analysed as an element constituting a nominal term because in this case, the NP included in the PP represent a new and different concept, like in (31).

(31) a. *mise à jour d<u>u</u>[de+le] site web*
'upgrade of the website'
b. *création d'<u>un</u> site web*
'creation of a website'

In this case, it is more useful to analyse the NP in (31a) having two different independent terms, namely *mise à jour* 'upgrade' and *site web* 'website'. Hence, determiners have the function of separating different terms.

Cadiot (1993) also agrees with this point of view: a PP without determiner designates a subclass of the preceding noun whereas the PP containing a determiner only describes the occurrence of the preceding noun. In the latter case, the classification can only be indirect: this results from the extensional property of the PP with determiner whereas in the former case, PPs can classify a noun on the basis of *intensional* properties.

(32) a. *chat à poils longs*
cat with hairs long
'longhaired cat'
b. *chat aux [à+les] poils mouillés*
'cat with wet hair'

The examples in (32) clearly show that (32a) without a determiner represent a subclass of cats whereas the version with a determiner (32b) describes a cat with wet hair, which is not a subclass of cats.

Anscombre (1990, 1991) adopt the same point of view when he states that PPs without determiner describe an essential propriety of the nominal head whereas PPs with determiner describe one of its accidental proprieties. He states that a property named P is an essential property of the entity named E if P can be considered as a unit that is inalienable of E. On the contrary, P is an accidental property if this property is temporary. Hence, an essential property is an inner property whereas an accidental property is only an actual state. The examples in (33) show that in the case of *bateau à voiles* 'ship with sails', which contains a PP without determiner, there are only a few adjectives that can be used to modify the noun after the preposition: these adjectives have to represent the type of the sail (33b) whereas in the version with determiner (33d), these have to designate the actual state of the sail.

(33) a. *bateau à voiles*
b. *bateau à voiles carrés/latines*
'ship with lateen or square-rigged sail'
c. *bateau à voiles ??hissés/*déchirés*
d. *bateau aux[à les] voiles hissés/déchirés*
'ship with hauled up/torn sails'

(Anscombre 1991, 26)

Cadiot (1993) observes that the situation is similar if the PP does not contain adjectives. The presence of the determiner imply that the element preceding or following the preposition designates an entity having an autonomous reference. This is confirmed by the following pairs:

(34) a. *un bagage à main*
'hand baggage'
b. *un bagage à la main*
'a baggage in the hand'

(35) a. *Jean a un bagage à main mais il le porte au[à+le] ventre.*
'Paul has a hand baggage but he is carrying it on his belly.'
b. **Jean a un bagage à la main mais il le porte au[à+le] ventre.*
'Paul has a baggage in his hands but he is carrying it on his belly.'

Anscombre (1991) states that a PP without a determiner cannot behave like a PP that designates a non evident property of the head. For example, a car intrinsically has a steering wheel but if a car functions with hydrogen, this is a non-evident property of the car.

(36) a. *voiture à *volant*
'car with steering wheel'
b. *voiture à hydrogène*
'hydrogen car'

(37) a. **un chat à deux oreilles*
'cat with two ears'
b. **vélo à roues*
'wheeled bike'

The examples in (37a,b) are not correct because of the same reasons: cats intrinsically have two ears and bikes have wheels – these are essential properties. If these complements are modified or extended, correct NPs can be obtained since a cat with three ears or a bike having a squared wheel are not evident.

(38) a. *un chat à trois oreilles*
'a three-eared sheep'
b. *un vélo à roues carrées*
'square-wheeled bike'

However, the statement of Cadiot (1993) according to which the function of a determiner is to introduce a new entity is not always true. In fact, there are terms that contain a determiner before the nominal component not representing a separate entity, like the examples in (39).

(39) a. *cancer de la peau*
'skin cancer'
b. *vidéo à la demande*
'video-on-demand'

It would be difficult to explain why the term in question is *cancer de la peau* instead of *cancer de peau*. In a previous study Nagy (2009) showed that the proportion of NPs with internal determiner is nearly 7% in comparison with the totality of nominal terms but the proportion of NPs with determiner that can also appear without determiner was not calculated. Consequently, completely aware of the loss that it represents, terms with determiners will not be considered as possible terms during the automatic extraction process.

In the example corpus, there were only ten cases where the PP complement of a term included a determiner, two of which are represented in (40) and (41):

(40) a. *reine des prés*
queenof+the meadows
'meadowsweet'

(41) b. *sensation de la faim*
sensation of the hunger
'sensation of hunger'

Another interesting characteristic of PPs in terms is the fact that prepositions can be left out (e.g. in (42)). This observation is mainly true for recently created terms on which Béjoint and Ahronian (2008) state that this omission is due to the effect of English. Nevertheless, the order of the nouns follows French rules.

(42) a. *code source*
'source code'
b. *accès Internet*
'Internet access'

(Béjoint & Ahronian 2008, 653)

This latter is not a problem for the automatic extractor because the pattern recognising nominal compounds will recognise them without modification.

5.3 Comparison of NPs and terminological noun phrases

In Sections 5.1 and 5.2, the main differences between PPs and APs in nominal terms and NPs were treated in detail. These sections were not based on articles on terminology extraction but mostly on linguistic articles. The differences concluded from the analysis are summarised in Table 1.

Table 1. Differences between nominal terms and NPs

	Nominal terms	NPs
APs	almost only postnominal (mainly relational and non-classifying adjectives)	pre- or postnominal (classifying, non-classifying, relational or ordinal adjectives)
PPs	almost without determiners, because a PP without determiner designates a subtype of the head	with or without determiners (designating a subtype or actual state)

6 Results

In this Section, I present the rule set which was elaborated on the basis of Chapter 4 and 5 and which were integrated into the term extractor. The second part of this section describes the efficiency of the term extractor in every phase of the term extraction process.

6.1 Rule set

The syntactic rules implemented into the term extractor are presented in a regular expression format, using the standard part-of-speech category abbreviations. This rule set covers most of the rules that were created on the basis of the analysis in chapter 5 but not all the rules that were used in the term extractor. Sign + indicates an occurrence of at least one, * indicates an occurrence of zero or more.

(1) N+
(2) N+ A* (Prep N A*)+
(3) N Prep V-INF

Rule (1) extracts one noun or a sequence of nouns, rule (2) nouns with prepositional complements and rule (3) nouns followed by a preposition and an infinitive. As it can be seen from the rules, neither prenominal adjectives nor PPs with determiners are taken into consideration.

6.2 Efficiency of the term extractor

In the field of computational linguistics, efficiency of an application is measured by three values: recall, precision and F-value. In order to calculate the effectiveness of the term extractor, I will use the same metrics. In term extraction, recall is the proportion of correctly extracted terms and of all the real terms in the corpus. Precision is the ratio of correctly extracted terms to all extracted terms. F-value is the harmonic mean of recall and precision.

$$recall = \frac{number\ of\ correctly\ extracted\ terms}{total\ number\ of\ real\ terms}$$

$$precision = \frac{number\ of\ correctly\ extracted\ terms}{total\ number\ of\ extracted\ terms}$$

$$F - value = \frac{2 * precision * recall}{precision + recall}$$

To provide a baseline, I also run the application with a list of rules that recognise all NPs, and I also executed two other applications using rule-based and/or statistical modules: these are Fastr (Jacquemin 2001) and YaTeA (Aubin and hamon 2006). These two extractors rely on the TreeTagger POS-tagging[9] program, so their relatively low metrics may be due to the fact as well. The other factor that influences their efficiency is that these term extractors were not created to extract terms specifically from French patent texts. The baseline values are represented in Table 2:

[9] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

Table 2. Baseline results (YaTeA, Fastr term extractors and my own application used with patterns recognising all NPs)

	G06F			A23L		
	recall	precision	F-value	recall	precision	F-value
YaTeA	0,5826	0,3045	0,3983	0,5711	0,3451	0,4270
Fastr	0,5349	0,3962	0,4523	0,5764	0,4130	0,4806
All NPs	0,5457	0,3072	0,3931	0,5615	0,3381	0,4221

Table 2 clearly shows that all the three present nearly the same values, thus they can be considered a real baseline to be compared to my own results.

These metrics were computed on the two different types of corpora (G06F as the IT corpus and A23L as the Human necessities corpus) in the case of my own application recognising nominal terms, as well. Firstly, these values were measured when the application did not use rule-based filtering, and secondly, when the term extractor was expanded by the filtering of stopwords. Table 3 shows the results of the term extraction process with or without filtering.

Table 3. The results of term extraction with or without filtering

Patent class	Rule-based filtering	Recall	Precision	F-value
G06F	No	0,8159	0,5847	0,6812
G06F	Yes	0,8311	0,6605	0,7360
A23L	No	0,7413	0,5664	0,6422
A23L	Yes	0,7599	0,6306	0,6892

As the results show, high recall can be achieved even without any filtering (0,82 in the G06F corpus, and 0,74 in the A23L): this is due to the fact that the structure of nominal terms complies with the preliminary patterns. However, the usage of syntactic patterns results in relatively low precision, because non terminological units also match these patterns.

Filtering (RBF) did not provoke a big increase in recall values, since they got higher only by 0,01. However, as hypothesised, stopword filtering significantly increased the precision in both corpora: this augmentation was nearly 0,07 in both corpus. This is due to the fact that filtering out words that cannot be part of terms exclude non terminological term candidates.

The used statistical methods (SF), the fine-tuning of which is yet a work to done, did not have the expected efficiency. The combination of the three used values, namely C-value, weight and weirdness, led to an overall increase of 0,01 of the F-value on both corpora (results not included in Table 3). A possibility of improving statistical results is the usage of machine learning algorithms; however, this technique requires a relatively large, annotated corpus where nominal terms are marked by hand.

7 Sources of error

Most of the problems with term extraction were due to incorrect part-of-speech tags associated to words. As POS-tagging was implemented by an automatic application, namely the Machinese of the Connexor company[10], these errors would not be easy to modify later on. Nearly 20% of the cases were caused by this source of error, that is 20% of the non-recognised terms (false negatives) and sequences incorrectly marked as terms (false positives) were due to the fact that at least one of the word in the sequence was associated with a wrong POS-code. For example, in the example corpus, *terminal* was tagged as an adjective 'final' instead of a noun 'terminal', and that was the same case with *anti-oxydant* that was often tagged as an adjective 'antioxidant' instead of a noun 'antioxidant'. An other frequent case was tagging a past participle as an adjective, like *utilisé* meaning 'used'.

Another frequent source of error was that the extracted candidate term was not really a term. It represented nearly 30% of the false positive cases. These non terminological units were for example *place* 'place' or *an* 'year'. In fact, these are the sequences which may be filtered out later on with the help of statistical measures.

Nearly 15% of the false positive and negative cases were provoked by the fact that in some cases, an AP or PP that are not part of a term were marked together with the nominal head as a term. These are exemplified in (43) and (44):

(43) *norme USB classique*
'classic USB standard'

(44) *préparation de crustacés*
'preparation of shellfish'

In (43) the AP *classique* 'classic' is not part of the nominal term *norme USB* 'USB standard', and in (44), the PP *de crustacés* 'shellfish' should not be included in the first term *préparation* 'preparation' but should be tagged as a separate term, *crustacé* 'shellfish'.

8 Conclusion

In this article, the internal structure of French NPs has been reviewed and has been compared with that of nominal terms with respect to their possible adjuncts and complements, with a specific stress on APs and PPs. This comparison was made in the purpose of establishing a nearly exhaustive list of different syntactic term structures in order that the term extractor recognise most of the possible terms.

Nominal terms do not have in general APs placed before the nominal head because specialised languages only admit classifying and relational adjectives that are placed after the noun. In certain cases, terms can have APs at their beginning, for example monosyllabic adjectives like *petit* 'little' or *long* 'long' or ordinal adjectives but if the latter ones precede the noun, they only designate an accidental quality of the noun, and consequently cannot be part of nominal terms. This is also proved by the results, because in the example corpus, only one case was found where the adjective preceded a noun.

[10] http://www.connexor.eu/technology/machinese/index.html

Complements of nominal terms in general cannot be PPs with a determiner since a determiner introduce a new entity, that is a new concept and terms can only represent one concept. Even if some terms can have PPs with a determiner, their proportion is really insignificant, hence it is not reasonable to consider them as being part of a term because that would result in too much noise during the extraction process. This was also confirmed by the results: there were only ten cases where the PP complement of the nominal head contained a determiner.

The most important message of the results is that term extraction can be efficient not only with the help of statistical methods but also with linguistic methods, especially on patent corpora and when recall values are more important. The rule-based term extraction provided high recall values with middle precision values, but the latter values could significantly be increased by rule-based filtering. At this time of the research, the chosen statistical methods (SF) only provoked a little augmentation in the average metric, the F-value.

References

Abeillé, Anne **&** Daniéle Godard. 1999. La position de l'adjectif en français: le poids des mots. *Recherches Linguistiques de Vincennes* 28. 9–32.

Ahmad, Khurshid, Lee Gillam **&** Lena Tostevin. 1999. University of surrey participation in trec8: Weirdness indexing for logical document extrapolation and retrieval (wilder). *The Eighth Text REtrieval Conference (TREC-8).*

Anscombre, Jean-Claude. 1990. Article zéro et structuration d'événements, In Michelle Charolles, Sophie Fischer & Jacques-Henri Jayez (eds.), *Le discours : représentations et interprétations*, 265–300. Nancy: PUN.

Anscombre, Jean-Claude. 1991. L'article zéro sous préposition, *Langue française* 91. 24–39.

Aubin, Sophie **&** Thierry Hamon. 2006. Improving term extraction with terminological resources, *Advances in Natural Language Processing Lecture Notes in Computer Science* 4139. 380-387.

Béjoint, Henri **&** Céline Ahronian. 2008. Les noms composés anglais et français du domaine d'Internet: une radiographie bilingue, *Meta : journal des traducteurs*, 53(3). 648–666.

Bosredon Bernard **&** Iréne Tamba. 1991. *Verre à pied, moule à gaufres* : préposition et noms composés de sous-classe. *Langue française* 91. 40–55.

Bouchard, Denis. 1998. The distribution and interpretation of adjectives in French: a consequence of Base Phrase Structure, *Probus* 10. 139–183.

Cabré, Maria Teresa. 1999. *Terminology. Theory, methods and applications*, Amsterdam/Philadelphia: John Benjamins.

Cabré, Maria Teresa, Rosa Estop'a Bagot **&** Jordi Vivaldi Palatresi. 2001. Automatic term detection. A review of current systems. In Didier Bourrigault, Christian Jacquemin **&** Marie-Claude L'Homme (eds.), *Recent advantages in Computational Terminology, 53-87*. Amsterdam/Philadelphia: John Benjamins Publishing Co.

Cadiot, Pierre. 1993. *À* entre deux noms: vers la composition nominale, *Lexique* 11. 193–240.

Cinque, Guglielmo. 1994. On the evidence for partial N-movement in the Romance DP, In Guglielmo Cinque (ed.), *Path Towards Universal Grammar. Studies in Honor of Richard Kayne*, 85–110. Washington: Georgetown University Press.

Frantzi, Katerina T. **&** Sophia Ananiadou. 1997. Automatic term recognition using contextual clues. In *Proceedings of Mulsaic 97, IJCAI*, Japan, 1997.

Jacquemin, Christian. 2001. *Spotting and discovering terms through natural language processing.* Cambridge(MA)/London: MIT Press.

Justeson, John S. **&** Slava M. Katz. 1995. Technical terminology: some linguistic properties and an algorithm for identification in text, *Natural Language Engineering*, 1(1). 9-27.

Laenzlinger, Christopher. 2003. *Initiation à la Syntaxe formelle du français: Le modèle Principes et Paramètres de la Grammaire Générative Transformationnelle*. Bern: Peter Lang AG.

Maynard, Diana & Sophia Ananiadou. 2000. Identifying Terms by their Family and Friends. *Proceedings of COLING 2000*, Luxembourg, 530-536.

Nagy, Ágoston. 2009. La structure interne des termes techniques du français et leur reconnaissance par ordinateur. In Anna Kieliszczyk & Ewa Pilecka (eds.), *La perspective interdisciplinaire des études françaises et francophones*, 117–123 Łask: Oficyna Wydawnicza LEKSEM.

Riegel, Martin, Jean-Christophe Pellat & René Rioul. 2009. *Grammaire méthodique du français* (4th edition) Paris: PUF.

Wüster, Eugen. 1976. La théorie générale de la terminologie, un domaine interdisciplinaire impliquant la linguistique, la logique, l'ontologie, l'informatique et la science des objets, *Actes du colloque international de terminologie, Québec 5–8 octobre 1975*, Québec: L'Éditeur officielle du Québec.