

## *Linguistic Elements of Terminological Units for Their Automatic Extraction*

R. Estopà

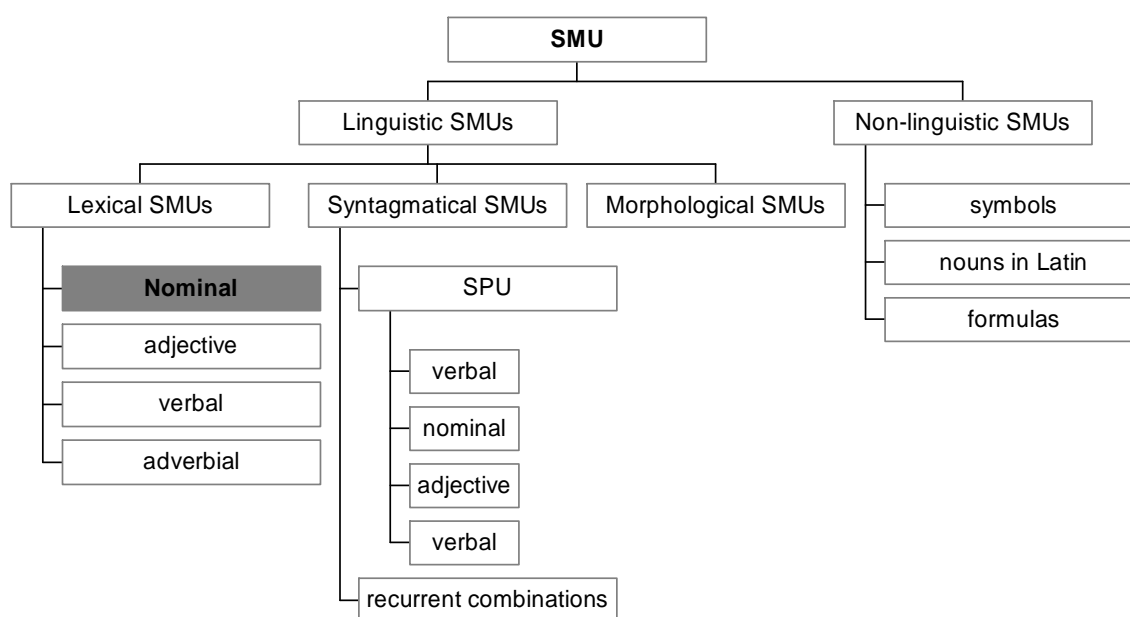
### *1. Objective*

The objective of this report is to disclose linguistic elements that may allow the construction of strategies to automatically identify and extract terminological units (TU) from specialized texts.

### *2. The Object of the Analysis*

We start from the premise that a TU is a specialized meaning unit (SMU) with a lexical character, referential capability, nominal category and specialized meaning in a concrete domain. *Brain, foot, fever, hepatitis, immunity, Alzheimer's disease, breast cancer, cerebral clot, SNC, TAC*, etc. are examples of medical TUs.

But, according to the criterion of specialists, TUs are not the only specialized units in specialized texts. There is a diverse wide range of specialized units according to the nature, structure and grammatical category of the different SMUs,<sup>1</sup> although we will give the priority to TUs in this case:<sup>2</sup>

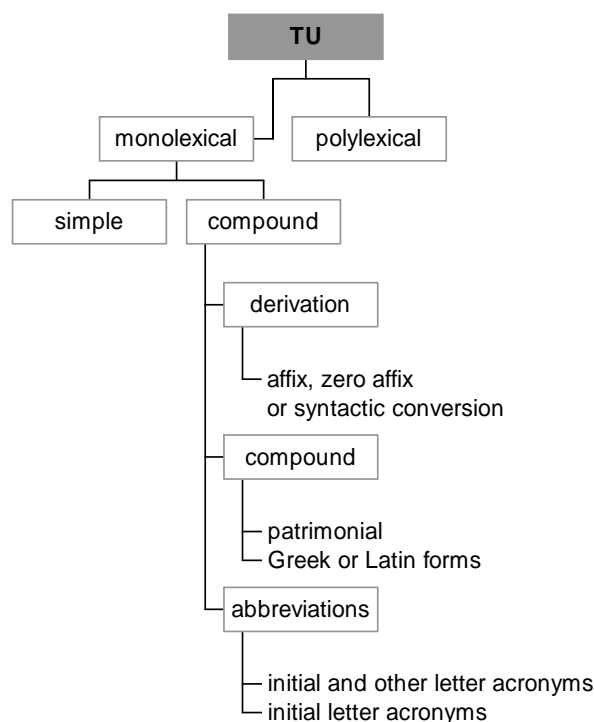


If we concentrate on TUs, we will find that they are also diverse. From the structural point of view, they can be morphological (*-itis, -oma, hepato-*), monolexical (*brain, histopathology*) or polylexical (*gastric neurosis, banda de Broca*). From the point of view of the way they are formed, monolexical units are at the same time divided in simple (*memory, axon, hernia, pain, hand*) and derived (*drain, treatment, injection, operation, diagnoratic*), compound (*appendicitis, anorexia, x ray, histopathology*), or acronyms (*PET, TAC, SNS, SNP*).

<sup>1</sup> SMU: specialized meaning unit; TU: terminological unit; SPU: specialized phraseological unit.

<sup>2</sup> We have neither included specialized morphological units (*-itis, -osis, -oma*, etc.) nor iconographic units.

Therefore, the object of the analysis of this study is the terminological unit, and the *Nominal Lexical units are the most prototypical terminological*, no matter what its morphological structure might be:



### 3. Basic Premises

Most of the terminological extractors have been centered in the retrieval of the polylexical terminological units through formal strategies based on its morphological-syntactical patterns. In this study, following the criterion of the specialist, we have the purpose of expanding the limits of the extraction of TU terminology and achieving more precise results. We start from the following premises:

1. Lexical units with a specialized meaning give priority to some morphological, morphological-syntactical, morphological-semantic, and pragmatic characteristics in relation to the lexical units with a general meaning.
2. The formal characteristics of TUs are not discriminative.
3. The application of strategies based exclusively on a sole (morphological, structural or morphological-semantic) parameter does not allow its automatic recognition.
4. The use of the same strategies to retrieve all the TUs is not efficient.
5. The application of strategies that are only based in the form of the units generates silence and noise.

As a consequence, we start from the hypothesis that we should find detection strategies that are based on a combination of elements with a basically linguistic character: lexical, morphological, morphological-syntactical, morphological-semantic elements, and also typographic, distributive and statistical ones. The combination of elements of a different kind is, then, the one that will allow us to recognize TUs automatically.

### 4. Corpus

We have based this study on the analysis of TUs from a medicine textual corpus in Catalan, which is formed by all the chapters about infectious diseases of Farreras' and

Rozman's book titled *Medicina interna* [Internal Medicine] (1997).<sup>3</sup> It is a specialized document written by specialists and destined both for medical doctors as well as for medical students. The corpus is made up by 61,000 occurrences. For the manual emptying of the contents, we have used the chapter section dealing with the diseases produced by rickets, with 12,069 occurrences.

## 5. Elements to Recognize the TUs

### 5.1 The Simple Monolexical TU

Simple monolexical TUs are difficult to treat automatically because their specialized character is totally idiosyncratic. They are, therefore, a kind of unit that have neither morphological nor explicit syntactical characteristics that allow their automatic detection. So, we must resort to lexical and/or contextual strategies to gain access to them.

The analysis of the lexicographical and textual corpus proves that the number of simple terms in relation to other specialized units is very low. However, it is interesting to be able to extract the simple TUs from texts because they are key pieces in specialized vocabulary. Firstly, because they are the term base formed by derivations. Secondly, because they are the nuclei of many syntagm units, which are the hyperonym or the meronym.

### *Automatic Extraction Strategies*

As a consequence, we believe that there are two strategies that might detect the simple monolexical SMUs. The first one is the use of a dictionary of nominal and verbal simple SMUs that are morphologically deployed in such a way that each entry can be associated with a semantic class label. The second one is the use of a more complex strategy based on the elements that its linguistic context provides about the SMUs: the linguistic resources of the text that indicate the presence of SMUs (metaterms, paraterms, generic terms, connectors, fixed patterns, typographic marks, etc.).

### 5.2 The Complex Monolexical TU: Derived, Compound and Acronyms

The complex monolexical TUs that a system would have to be able to extract automatically from texts are derived units, patrimonial compounds, neo-classical compounds and acronyms.

#### 5.2.1 Derived TUs

From the lexicological-morphological point of view, derived TUs present three singular characteristics. In the first place, they are formed from a lexical base that is always a specialized meaning unit. In the second place, these bases combine with a much more limited number of affixes than the affix attached to the non-specialized lexical bases. And, in the third place, the lexical base of derived TUs is usually related to a simple meaning unit or with a pertinent classic form within the medical domain (In Catalan: *postinfart, prepart, subencèfal, circulació, desintoxicació, tòxic, destil·lació, infiltració, infiltrar, intubació, sensibilitat, immunitat, toxicitat, encefàlic, neurològic, immunològic, endotelial*). According to these three observations, it seems indispensable that an extractor should associate units that share the same lexical base and relate them with a simple SMU and/or a Greek and Latin form.

---

<sup>3</sup> Farreras, P. and C. Rozman (1997). *Medicina interna*. Madrid: Harcourt Brace (13<sup>th</sup> edition).

### *Strategies for Automatic Extraction*

Based on the characteristics of derived SMUs, the strategy that we propose to detect the pertinent derived monolexical SMUs is, firstly, to group the units in the text that present the same lexical base and, later on, relate this lexical base with a simple unit or with a pertinent Greek or Latin form in the medical domain (in Catalan):

*orofaringi, -íngia, orofaringe, faringe* ⇒ *oro-, faringe*<sup>4</sup>  
*epidèrnia, epidermis, dermis,* ⇒ *epi-*<sup>5</sup>, *-dermi*<sup>6</sup>  
*histologia, histològic, -a, histològicament* ⇒ *histo-*

To carry out this process, a dictionary of Greek or Latin forms and a dictionary of simple SMUs would be needed to act as filters.

#### *5.2.2 Patrimonial Compounds*

Patrimonial compounds are not productive in the medical domain. They are even less productive in texts written for specialists. The few nouns formed by words in the Catalan language are used in oral speech or in public information texts.

Through the analysis of the corpus and medical dictionaries, we have seen that this kind of nouns name instruments (in Catalan: *comptagotes, comptaglòbuls, abaixallengües, tirapits, tirallet, portaagulles, portacames, portacuixes, portadrenatge, portaplaquetes*) or medicinal plants (in Catalan: *matafaluga, matafoc, matallums, mataparent, matapoll*).

We have also seen that all the patronymic compounds we have documented present the same morphological-syntactical structure: [V[N]<sub>SN</sub>]<sub>N</sub>; that is, a verbal nucleus and a noun that is the thematic argument of the verb. Besides, we have verified that, in such a case that the compound refers to an instrument, the noun of that construction is always a term of the same specialized topic (In Catalan: *tirapits, abaixallengües, portaplaquetes*, etc.). Semantically speaking, the names of the instruments that respond to this structure may be described as “*an instrument used for V<sub>function</sub> + N<sub>object</sub>*”. Therefore, they are formed by an exocentric nucleus, and its predicate may be interpreted literally from the meaning of its components. However, the meaning of the names of plants is not literal, but always of a metaphorical order.

#### *Automatic Extraction Strategies*

To detect patrimonial compounds, an extractor may relate the second component with simple terms from the dictionary or with the derivatives from the corpus and test if they have a specialized character in the medical field of reference. This is the case of the compound names of instruments in which we have verified that the second constituent is always a TU (in Catalan): *abaixallengües, comptaglòbuls, comptagotes, tirapits, portaagulles, portacuixes, portadrenatge*, and *portaplaquetes*.

In relation to the vulgar names of plants, the strategy to relate the lexical bases is not valid because they are bases that are not literally specialized since they are the result of a metaphoric process that the extractor cannot break down. In these cases, we must resort to a more idiosyncratic way such as a dictionary.

#### *5.2.3 Neo-classical compounds*

---

<sup>4</sup> And the Catalan terms integrated by the form *oro-* would also be related: *orocinasa, orolingual, oronasal*, etc.

<sup>5</sup> And also the Catalan words that are formed by the form *epi-*: *epidèmia, epicauma, epigastri*, etc.

<sup>6</sup> And also the Catalan *gerodèrnia, helodèrnia, crisodèrnia, espasmodèrnia*, etc.

The results of the analysis of our corpus allow us to confirm that neo-classical compounds are very numerous in medicine and that they are the bases of most of its specialized vocabulary. As it has been said by López Piñero and Terrada Ferrandis (1990), about a thousand Greek or Latin roots and about eighty classic affixes make up almost all the vocabulary of medicine. This means that, in health sciences, a relatively small group of Greek and Latin elements generate a very high number of units.

Besides, it is interesting to consider that health care professionals do not know the whole lot of medical words, but that they are capable of deciphering their meaning without ever having even seen or heard them before. And they are able of doing so because they know the formation mechanisms of most of the medical words, and have internalized a group of Greek and Latin forms that allow them to code and decode the meanings of the words that are being used in medicine.

#### *Strategies for Automatic Extraction*

To extract the neo-classical compounds, according to the logic that characterizes the knowledge a specialist has of medical terminology, an extractor can simulate the strategy of the professional by using a dictionary with 1100 pertinent Greek and Latin forms in the medical domain and a reduced number of rules to combine the forms.

But besides the usefulness of the classic forms for the detection of the SMUs that are compounded in the cultured way, these may serve, in the first place, to test if one lexical base is especially pertinent or not. So, the detection of Greek or Latin forms allows us to establish an identification chain that begins with the detection of the Greek or Latin form and finishes with the identification of the polylexical SMUs (in Catalan):

Cultured forms  $\Rightarrow$  derived and/or compound  $\Rightarrow$  syntagm units

hepat(o)- + -itis  $\Rightarrow$  *hepatitis*  $\Rightarrow$  *hepatitis vírica*, *hepatitis crònica agressiva*, *hepatitis epidèmica*, *hepatitis fulminant*, *hepatitis supurada*, etc.

sero- (*ser-*), -logia, sèrum, serologia, serològic, -a, serològicament, seroaglutinació, seromucós, serologia, sèrum d'Hayem, sèrum hemàtic, sèrum polivalent<sup>7</sup>, etc.

In the second place, the possibility to relate all the SMUs that share the same root with a simple term or with a classic form, not only makes automatic extraction easier, but also helps the user to select the pertinent units from those proposed by the system because the productivity of a particular morphological-semantic root in medical texts is one more element to decide if a unit is specialized or not.

Consistent with this multifaceted character, it would be both useful and productive if an extractor could construct series of words that would share at least one root or a pertinent form through a dictionary of Greek and Latin forms and a dictionary of simple terms.

#### *5.2.4 Acronyms*

Acronyms with specialized meanings are very much present in specialized texts, especially in those intended for specialists. Acronyms (the combination of the initial letters of a polylexical SMU) apparently are simple lexical units but, if we analyze their formation process, we can see that they have a complex origin. Acronyms present characteristics that single them out:

1. They frequently have an international character.

---

<sup>7</sup> In the *Diccionari Enciclopèdic de Medicina* (DEM) (1990) there are 203 meanings for *sèrum*.

2. They are semantically opaque since, formally, the only relationship they keep with the syntagm they substitute is the graphic form of their initials.
3. They present a singular typographic form since they are usually written in capital letters, normally close together with no periods or blank spaces in between.<sup>8</sup>
4. They are usually formed by combinations of two to five letters that, normally, substitute a segment of three referential words.<sup>9</sup>
5. They usually appear in the text written inside parenthesis following the whole-developed segment they substitute.<sup>10</sup>

### *Extraction Strategies*

Starting from their linguistic characteristics, extractors have two possible ways to detect acronyms in a text:

- by using an acronym dictionary
- by resorting to some aspects in the text (the location of the acronyms in the text, typography, determination, etc.)

The first option presents two inconveniences. One, having a dictionary of medical acronyms in Catalan. A dictionary that—if it is ever compiled—would have to be very extensive. An English dictionary of medical acronyms was published in 1989 and it already included more than 15,000 [Heister, 1989]). And, two, using an acronym dictionary would not allow the detection of new acronyms, which would be an important problem in a professional domain, considering the high level of lexical innovation we find in health sciences.

The second way for acronym detection seems quicker from the linguistic and computer science points of view. It implies resorting to pragmatic-linguistic heuristics that may take advantage of the characteristic elements of acronyms we have just made comments about. Our option is inclined in favor of a mixed strategy in which both kinds of resources—the dictionary and heuristics—become complementary.

### *5.3 The Polylexical Terminological Units (PTU)*

Generally speaking, the traditional terminology extractor detects all the explicit PTUs in specialized texts through a group of structural patterns. The exhaustive character of the identification of this kind of units is very high since they just do not detect the implicit units in the text. That is, the overlapping units that include more than one TU and, above all, the *units hidden* by a discursive anaphora. However, to favor search parameters (the corpus and the final purpose of the emptying of the contents), not to retrieve these units is not a problem because, frequently, if a lexical unit appears as an anaphora or is overlapped in one place of the text, this same unit is explicit in another place of the corpus.

So, the results collected from the analysis by the main automatic extractors allow us to say that, in general, the retrieval of the PTUs does not generate silence, but, instead, they generate much noise because the morphological-syntactical structures that these systems use to extract the PTUs are not exclusive of these units. It has been estimated that between 45% to 75% of the candidates proposed by these systems have to be

---

<sup>8</sup> If an acronym is written in small letters, that is a sign of its high degree of lexicalization—as in laser, radar, aids—, in which they are treated as simple units.

<sup>9</sup> This seems logical, if we think that there are no TUs over five referential elements, and that the majority are formed by a noun and one or two complements.

<sup>10</sup> In the case of acronyms that are very well known to specialists, they do not appear inside parenthesis and this loss of their relation with the syntagm they substitute is the main problem extractors find, since an isolated acronym is idiosyncratic.

refused. This leads us to conclude that the identification strategy of the PTUs that are only based on morphological-syntactical patterns is a too permissive filter.

To be more restrictive it is necessary to propose other noise-reduction mechanisms to filter the discursive segments that present analogous structures. Some authors have proposed strategies in this way, to complement structural filters: statistical strategies (Bourigault 1993), syntactical strategies (Jacquemin 1994), semantic strategies (Naulleau 1998) and contextual strategies (Pearson 1998). Our proposal is to use strategies of a different nature that, combined, would first eliminate the discursive units, and then make the distinction between PTUs, SPUs and recurrent combinations easier.

Like most extraction systems, we have centered the detection of PTUs in the two most productive structures, which are the base of all the polylexical SMUs in medical texts, and in general in all the specialized texts in romance languages; that is,  $[N[A]_{SA_{Adj}}]_{SN}$  and  $[N[\text{of}(\text{art})][N]_{SPrep}]_{SN}$ .

### 5.3.1 $[N[A]_{SA_{Adj}}]_{SN}$

In Catalan and in Spanish, the structure  $[N[A]_{SA_{Adj}}]_{SN}$  is the most productive of the specialized texts, but, at the same time, it is one of the most noise-producing ones, since superficially, it cannot be known if it is specialized or not. The analysis of the medicine textual corpus shows that the specialized or discursive character of a unit with a  $[N[A]_{SA_{Adj}}]_{SN}$  structure depends on the specialized nature or not of the noun or adjective that integrate it. Following this line, we distinguish four possibilities:

$$\begin{aligned} [N_{\text{esp}} [A_{\text{esp}}]_{SA_{Adj}}]_{SN} &= \text{PTU} \\ [N_{\text{no esp}} [A_{\text{esp}}]_{SA_{Adj}}]_{SN} &= \text{PTU} \\ [N_{\text{esp}} [A_{\text{no esp}}]_{SA_{Adj}}]_{SN} &= \text{PTU or discursive unit (DU)} \\ [N_{\text{no esp}} [A_{\text{no esp}}]_{SA_{Adj}}]_{SN} &= \text{DU or lexical unit (LU)} \end{aligned}$$

From these possibilities, we have concluded the following results:

- a) If the nucleus of a unit is a SMU and the adjective that complements it is also a SMU, the result is always a PTU (in Catalan):

*antibiòtic bactericida, coll uterí, degeneració lenticular, dermatitis actínica, embòlia cerebral, inhibició enzimàtica, injecció endodèrmica, intervenció quirúrgica, llengua leucoplàstica neurosi gàstrica, punció abdominal, tractament mèdic, etc.*

- b) If the noun of the unit is not a SMU, but the adjective that classifies it is a SMU, the resulting combination is also a PTU (in Catalan):<sup>11</sup>

*bossa serosa, canal coclear, capacitat pulmonar, punt alveolar, timbre nasal, vas limfàtic, etc.*

---

<sup>11</sup> This is a very productive combination when the PTU belongs to the semantic class of the parts of the human body, but the nucleus of the unit refers to a part of the body that receives the name of an object of the material world with which it has a certain degree of similarity.

- c) However, if the noun is a SPU and the adjective is not, it might be either of two kinds of units. It is either a PTU (in Catalan: *augment alt, aplicació ràpida, berruga plana, desaparició ràpida, difusió mundial, disminució lenta, edema maligne, eritema simple, febre groga, laringitis aguda, nefritis local, població nova*, etc.) or else a DU with no specialized character, (in Catalan: *hepatitis rara, radiografia dolenta, injecció forta, pacient diferent, limfopatia generalitzada, reacció adversa, reacció local, tractament actiu*, etc.). In this second case, the system would have to propose the nominal nuclei as terminological and get rid of the noise generated by the adjective that modifies them.
- d) Finally, if a segment is not formed by any SMU, the resulting unit is either a DU (in Catalan: *element clau, estudi recent, mecanisme possible, viatger internacional*) or a lexical unit that is not pertinent to that specialized topic (in Catalan: *animal domèstic, conca mediterrània, continent americà, zona urbana*) and, therefore, there is no terminological interest in retrieving them from the texts in any one of the two cases.

### 5.3.2 [N[of (art) [N]<sub>SPrep</sub>]<sub>SN</sub>

Standing beside the [N[A]<sub>SAdj</sub>]<sub>SN</sub> structure, the [N[of (art) [N]<sub>SPrep</sub>]<sub>SN</sub> structure is the one that causes the highest noise indexes, even though it is not as productive as the first. This structure, on account of the characteristics presented by the complement (a frequently determinative prepositional syntagm), has less lexicalizing strength than the form. For that reason, we also find DUs, SPUs and recurrent specialized combinations together with the PTUs of this structure in specialized texts.

The pertinence of a unit with a [N[of (art) [N]<sub>SPrep</sub>]<sub>SN</sub> structure depends, as in the previous case, on the specialized nature or not of its constituents and, in some cases, also on the eventful character of the unit's nucleus and on the fact that the noun nucleus of the prepositional syntagm is underrated as proper or common. According to this premise, we distinguish the following combinations:

[N <sub>esp</sub> [of [N <sub>patronímic</sub> ]] <sub>Sprep</sub> ] <sub>SN</sub>	= PTU
[N <sub>esp</sub> [of (art) [N <sub>no esp</sub> ]] <sub>Sprep</sub> ] <sub>SN</sub>	= PTU or DU
[N <sub>esp</sub> [of (art) [N <sub>esp</sub> ]] <sub>Sprep</sub> ] <sub>SN</sub>	= PTU or recurrent combination
[N <sub>verbal noun</sub> [of (art) [N <sub>esp</sub> ]] <sub>Sprep</sub> ] <sub>SN</sub>	= SPU
[N <sub>verbal noun</sub> [of (art) [N <sub>no esp</sub> ]] <sub>Sprep</sub> ] <sub>SN</sub>	= DU
[N <sub>quant</sub> [of (art) [N]] <sub>Sprep</sub> ] <sub>SN</sub>	= DU
[N <sub>paraterm</sub> [of (art) [N]] <sub>Sprep</sub> ] <sub>SN</sub>	= DU
[N <sub>no esp</sub> [of (art) [N <sub>no esp</sub> ]] <sub>Sprep</sub> ] <sub>SN</sub>	= DU or LU
[N <sub>no esp</sub> [of (art) [N <sub>esp</sub> ]] <sub>Sprep</sub> ] <sub>SN</sub>	= PTU

This table of combinations allows us to formulate the following conclusions:

- a) The [N[of (art) [N]<sub>SPrep</sub>]<sub>SN</sub> structure results in a PTU and, in medicine, is the most frequent structure with a prepositional complement (in Catalan):<sup>12</sup> *amígdala de*

<sup>12</sup> From the data, we have verified that the proper noun that specifies a term always appears indeterminate because, in Catalan and in Spanish, indetermination reinforces the generic character of the unit. Semantically, the sequence can always be expressed in the following paraphrases:

*Y ha descubierto o ha padecido X, en el que Y es un investigador o un paciente famoso o el primer paciente y X, la cosa descubierta o padecida.*

*Luschka, banda de Broca, distròfia de Fuchs, flegmó de Monat, histerectomia d'Amman, malaltia de Miller, membrana de Browman, òrgan de Corti, tendó d'Aquiles, etc.*

- b) If the first noun in the sequence [N[of (art) [N]<sub>SPrep</sub>]<sub>SN</sub> is a term, but the second is not, then the resulting combination may be a DU (in Catalan): *zoonosi de distribució mundial, xeringa de l'agulla llarga, etc.* (although this discursive combination is not very frequent in the texts written by specialists and destined for specialists) or a PTU (in Catalan): *filtre de vidre, crani de torre, fetge d'acordiò, febre dels pantans, etc.*
- c) If the two nouns of the sequence are terminological, according to the semantic characteristics of their constituents, we have either a PTU (in Catalan: *hipoglucèmia d'esforç, síndrome d'immunodeficiència, etc.*) or a specialized combination (in Catalan: *paràlisi del nervi bucal, paràlisi del braç esquerre, picada de paparra, etc.*). We have verified that, normally, there is an absence of the article in the first case and that the article is usually present in the second case.
- d) The examination of the textual corpus and specialized dictionaries shows that one of the most productive combinations in medicine is the one constructed with a verbal noun and a prepositional syntagm introduced by the preposition *de* (of), that has a nominal term as a nucleus, although this sequence is always a SPU (in Catalan):

*absorció del sèrum, acumulació de líquid intravascular, augment de la permeabilitat vascular, elevació de la creatinafosfocinasa, instauració del tractament, prevenció de la malaltia, tractament de la febre botonosa, etc.*

We have also verified that the prepositional syntagm in most of the nominal SPUs that present this structure is determined by the definite article.

- e) If we find a [N[of (art) [N]<sub>SPrep</sub>]<sub>SN</sub> sequence in which the base of the first noun is a non-specialized verb and the second noun is not terminological, we may conclude that the sequence is discursive (in Catalan): *augment del nombre de factors, eliminació de la vegetació, etc.* Data prove, however, that this combination is not very frequent in specialized medical texts.
- f) However, we have very frequently found polylexical sequences in specialized texts that have a quantitative noun as a nucleus (in Catalan: *majoria de casos, la resta de símptomes, la totalitat dels pacients, el conjunt de característiques*) or a noun that makes the structuring of speech easier (in Catalan: *(al) final del capítol, (al) llarg de l'article, (a l'inici del capítol, etc.*<sup>13</sup>) and that are always discursive complex units. In the first case, only the structure's nucleus rouses noise, because the complement tends to be a TU. However, in the second case, we always have to refuse the whole segment.
- g) If the first noun of a [N[of (art) [N]<sub>SPrep</sub>]<sub>SN</sub> sequence is a paraterm (that is, a word that precedes a term and that implicitly gives semantic information about it), the noun syntagm of the complement of these sequences is always a SPU (in Catalan):<sup>14</sup>

---

[Y has discovered or has suffered X, in which Y is a researcher or a famous patient or the first patient and X the discovered or suffered thing.]

In all the appearances presented by this structure, the nucleus and the PTU belong to one of the following conceptual classes: diseases or pathological states; treatments; methods and techniques, parts or components of the human body; instruments; actions and operations.

<sup>13</sup> Notice that these kinds of nouns are frequently preceded by the locative preposition *a*.

<sup>14</sup> Paraterms are very abundant in specialized texts and, even though they produce noise, they are semantically interesting because they provide pragmatic (topographic, metalinguistic, temporal,

*causa del xarampió, presència de taca negra, denominació de rickètsia, començament de la malaltia, característica de la malaltia, existència d'infeccions subclíniques, etc.*

- h) If none of the constituents in a [N[of (art) [N]<sub>Sprep</sub>]<sub>SN</sub> structure is a term, it is a segment with no specialized value that generates noise and can either be a DU (in Catalan: *mes de maig, darrera dels anys vuitanta*) or else a lexical unit that is not pertinent in medicine (*bossa de pa, estació de l'any, rellogge de sorra*, etc.)
- i) Finally, if we find sequences with a [N[of (art) [N]<sub>Sprep</sub>]<sub>SN</sub> structure in medical specialized texts, whose first noun is not terminological, but the second is, and, besides the first noun is neither a paraterm, nor a quantitative, nor a discursive structure organizer, it is a TU.

Data show that, in such cases, the resulting unit usually names parts of the human body or portions of parts of the human body in which the complement is determined, because this determination underlines the unique character of that part in the resulting unit (in Catalan):

*ala de la ròtula, angle de la costella, apèndix del testicle, base del cor, escorça del cristal·lí, falç del cervell, istme de la pròstata, radi del cristal·lí, tenda del cerebel, vàlvula de l'urèter, vel del paladar, etc.*

#### *Strategies for Automatic Extraction*

After this brief analysis of the TUs, we believe that, through strategies that are based on lexical, morphological or morphological-syntactical elements, a system may identify all the TUs presented, except the following three combinations, in which it is not possible, with only formal criteria, to know if it is a PTU, a UD or a recurrent combination:

[N<sub>term</sub> [A<sub>no esp</sub>]<sub>SAdj</sub>]<sub>SN</sub> = PTU or DU  
 [N<sub>term</sub> [de (art) [N<sub>no esp</sub>]<sub>Sprep</sub>]<sub>SN</sub> = PTU or DU  
 [N<sub>term</sub> [de (art) [N<sub>esp</sub>]<sub>Sprep</sub>]<sub>SN</sub> = PTU or recurrent combination

In these three cases, that lead to ambiguities if formal resources are used, it is indispensable to resort to semantics. So, to remove ambiguity from these combinations, we propose that an automatic extractor should resort to the use of morphological-semantic filters that consider the semantic class of the nucleus and complement of the polylexical unit.

Summarizing, in this study we have proposed elements with the basically linguistic character of the terminological units, and strategies based on those characteristics that an extractor could use to detect them automatically from texts, surpassing the results of classical extractors that are fundamentally based on their form.

Finally, we present a chart that synthesizes the strategies based on the linguistic characteristics of TUs that an extractor could use to detect them and the tools it would use in each case:

### **1. Detection of linguistic SPUs**

#### 1.1 monolexical

##### 1.1.1 simple:

##### **☞ dictionary of simple medical SPUs**

##### 1.1.2 derived:

---

relationship, morphological, etc.) information about TUs, and establish conceptual relationships between the terms in a text.

- ☛ **dictionary of simple medical SPUs**
- ☛ **dictionary of Greek or Latin forms**
- ☛ **protocol of root-family relationships**

1.1.3 patrimonial compounds

- ☛ **dictionary of simple medical SPUs**
- ☛ **protocol of root-family relationships**

1.1.4 neo-classical compounds

- ☛ **dictionary of Greek or Latin forms**

1.1.5 acronyms

- ☛ **dictionary of frequent acronyms**
- ☛ **instructions for acronym detection**

1.2 polylexical

1.2.1 PTUs

- ☛ **dictionary of simple medical PTUs**
- ☛ **dictionary of Greek or Latin forms**
- ☛ **dictionary of quantitatives and discursive organizers**
- ☛ **dictionary of medical paraterms**
- ☛ **structure filters**
- ☛ **semantical and formal filters for the NA and N of (art) N**

**structure**

- ☛ **protocol of root-family relationships**

1.2.2 SPUs

- ☛ **dictionary of simple medical SPUs**
- ☛ **dictionary of Greek or Latin forms**
- ☛ **dictionary of quantitatives and discursive organizers**
- ☛ **dictionary of medical paraterms**
- ☛ **structure filters**
- ☛ **semantical and formal filters for the N of (art) N structure**
- ☛ **tool to extract use frequencies**