Extracting Multiword Expressions using Enumerations of Noun Phrases in Specialized Domains: first experiences

Merley da S. Conrado[1], Walter Koza[2], Josuka Díaz-Labrador[3], Joseba K. Abaitua[3], Solange O. Rezende[1], Thiago A. S. Pardo[1], and Zulema Solana[2]

Universidad de São Paulo, Universidad Nacional de Rosario - CONICET, Universidad de Deusto
{merleyc,solange,taspardo}@icmc.usp.br,kozawalter@opendeusto.es
{josuka,joseba.abaitua}@deusto.es,zsolana@arnet.com.ar

Abstract. We present a recognition algorithm for enumerations of Noun Phrase (NPEs) whose objective is to detect and extract multiword expression (MWE). The algorithm used syntactic rules elaboration from linguistic information aiming to recognize NPEs. This information corresponds to morphological categories (noun, adjective, female, male, etc.). The evaluation takes into account only bigrams found in two different domain corpora of medicine and legal texts. The results are encouraging because, despite the low recall of MWEs, many significant terminological units from the two specialized domains were detected and extracted.

Keywords: Multiword expressions, enumerations of noun phrase.

1 Introduction

Multiword expressions (MWEs) are considered "*lexical items that: (a) can be decomposed into multiple lexemes; and (b) display lexical, syntactic, semantic, pragmatic and/or statistical idiomaticity*" [18]. Accordingly, their identification will benefit from expertise at the same linguistic levels.

Furthermore, MWEs play an increasingly relevant role in practical applications of natural language. Jackendoff [10] estimates that the number of MWE is similar to magnitude as the number of simple words in a speaker lexicon. In the case of specialized domains their significance is even higher because they normally correspond to the key concepts of a specific field. Several analytic techniques, including both symbolic and statistical methods, are being used to identify and collect MWEs to make them part of the dictionaries and thesauri of natural language processing systems. It has been also argued that techniques for extraction of multiword expressions would considerably benefit from deeper syntactic analysis [18].

Moreover, syntactic expressions may help in the MWE identification. Nenadic and Ananiadou [12] give examples of such expressions: "*an enumeration of terms (e.g., steroid receptors, such as, estrogen receptor, glucocorticoid receptor, and* *progesterone receptor), term coordination (e.g., adrenal glands and gonads), or conjunction of terms (e.g., estrogen receptor and progesterone receptor)*".

We depart from the hypothesis that enumerations, as other specific constructions (such as topicalizations, appositions, etc.) are often introduced to highlight the relevance of a given element of discourse. Furthermore, in the context of specialized domains these highlighted elements may correspond to terminological word combinations. There is an example of enumeration in the following text fragment (that it was adapted from the Wikipedia page about *derecho* - law).

Algunas de las diversas ramas jurídicas son derecho administrativo, derecho ambiental, derecho civil, derecho familiar y derecho religioso.
(Some of the various law branches are administrative law, environmental law, civil law, family law, and religious law.)

We also believe that the recognition of enumerations of noun phrase (NPEs) as multiword expressions will also render the extraction of complex terminological units. So, our hypothesis is that by identifying enumerations of noun phrase is possible to recognize terms of a specific domain. In this paper we describe a MWE recognition algorithm that is fed with candidates from a syntactic parser enriched with rules for the detection of enumerations of noun phrase. As we are going to see, NPEs are by and large good and successful MWEs candidates.

The target of our MWE extraction method is the identification of enumerations of noun phrase. We decided to experiment with NPEs because we believed these are a type of noun phrase (NP) that usually contain lexical items of a particular importance, which very often also are MWEs. We used corpora of two different domains: medicine and law. The results were evaluated by experts of each domain and we calculated their accuracy. These results were encouraging and showed that the syntactic method for the recognition of enumerations of noun phrase meets the expectations of MWE extraction. Most of these MWEs were also terminological units, containing the main concepts of the specialized domain; so their recognition has more relevance.

2 About Enumerations

An enumeration is a successive sequence of elements of the same class and that have the same syntactic function. In Spanish, each component is separated by a comma and, before the last component, normally a conjunction is introduced (1). However, there may be more elaborate sequences that include a comma as part of a complex expression, thus triggering the use of semicolons to make more visible the enumeration (2).

(1) [*Juan come manzanas, naranjas y fresas.*] (John eats apples, oranges and strawberries.)

(2) [*Los alumnos destacados son: Santamaría, Carlos; Taborda, Elena; Varlotta, José ...*] (The outstanding students are: Santamaría, Carlos; Taborda, Elena; Varlotta, José ...)

In case of complete enumerations, the last element has to be introduced by a conjunction and not by comma:

(3) [*Compró pan, verduras y carne.*] (He bought bread, vegetables and meat.)

(4) [*No compró pan, verduras ni frutas.*] (He did not buy bread, vegetables or fruits.)

(5) [*¿Compró pan, verduras o frutas?*] (Did he buy bread, vegetables or fruits?)

2.1 Enumerations classification

The literature on enumerations of noun phrases in Spanish is rather limited, and in our work we have followed Koza's classification [11]. This author provides an overview of enumerations based on two main aspects: (i) the structure of the components, which is, the relation among the elements in the sentence (depending on their structure, or their semantic features); and (ii) the number of components. Here we present the Koza' classification.

According to the structure of the components

The relationship among enumeration components is seen here. All enumerated elements must belong to the same syntactic category, i.e. we cannot put verbs and adjectives in the same enumeration. There are cases in that the components have the same structure, then we may identify simple enumerations, which are conformed by phrases (noun, adjectival, verbal, prepositional, and adverbial phrases), clausal enumerations, and mixed enumerations

Nevertheless, also there are enumerations with components of different categories, but with a similar syntactic function. E.g. in the following sentence:

(6) [*Llegó en su coche blanco, de llantas cromadas y que había ganado en un concurso televisivo.*] (He arrived in his white car, of chrome rims and that he had won on a game show.)

We have an enumeration conformed by an adjective ("blanco"), a prepositional phrase ("de llantas cromadas"), and a subordinate clause ("que se había ganado en un concurso televisivo") that are the noun "coche" modifiers. Koza denominates that kind of enumeration as "mixed" and they may be composed of phrases of different class or phrases and subordinate clauses.

In the following items, we show the enumeration classes according to the components features.

- *Phrasal enumerations*

Phrasal enumerations are divided into five groups: noun (7), adjectival (8), verbal (9), prepositional (10), and adverbial (11) enumeration. Each phrase could be conformed only by its heads or by its heads and complements.

(7) [*Compró pan, carne, agua mineral y verduras.*] (He bought bread, meat, mineral water, and vegetables.)

(8) [*María es alta, un poco delgada, rubia y muy linda.*] (Mary is tall, a Little thin, blond, and very beautiful.)

(9) [*Julián llegó, cocinó, comió opíparamente y descansó.*] (Julian arrived, cooked, ate sumptuously, and taked a rest.)

(10) [*Esta casa es de mi mamá, de mi papá y de mi hermano.*] (This house is my mom, my dad, and my brother.)

(11) [*Lo necesito aquí, ahora e inmediatamente.*] (I need you here, now and immediately.)

- *Clausal enumerations*

Clausal enumerations are divided into two groups:

Type I: Enumerations conformed by coordinated clauses which do not set a relation subordinated clause-superordinate clause (12);

(12) [*Juan canta, María baila y Agustín lee.*] (John sings, Mary dances, and Augustine reads.)

Type II: Enumerations conformed by subordinate clauses (13).

(13) [*Quiero que me comprendan, que no me juzguen y que me esperen.*] (I want that they understand me, which they do not judge me, and that they wait for me.)

- *Mixed enumerations*

Finally, mixed enumerations are divided into two groups:

Type I: Phrases combination:

(14) [*María es rubia, alta y de pelo largo.*] (Mary is blond, tall, and long hears.)

Type II: Phrases and subordinate clauses combination:

(15) [*Quiero una mujer de buenos modales, agradable, que me quiera y que sepa cocinar.*] (I want a nice woman, of good manners, which she loves me, and that she knows how to cook.]

According to the number of the components

Koza distinguishes two kinds of enumerations here:

- *Complete enumerations*

The complete enumerations are those that closing the list and not allowing the inclusion of new components. In this group, we may find the following complete enumerations:

a) Enumerations with final copulative and/or disjunctive conjunction.

(16) [*Platero es pequeño, peludo y suave.*] (Platero is small, hairy, and soft.)

(17) [*¿Platero es pequeño, peludo o suave?*] (Is Platero small, hairy, or soft?)

b) Enumerations with asyndeton

The asyndeton is a special case where there is no conjunction between the penultimate and the last component.

(18) [*Platero es pequeño, peludo, suave.*] (Juan Ramón Jiménez) (Platero is small, hairy, soft.)

– *Infinite enumerations*

Enumerations that do not have an explicit end and suggest other components are included in this group.

I- Enumerations that end with "etcetera"

(19) [*Le gustan todas las frutas; las manzanas, las naranjas, las sandías, etcétera.*] (He likes all fruits: apples, oranges, water melons, etc.)

II- Enumerations that end with the three period ellipsis

(20) [*Los números primos son el dos, el tres, el cinto, el siete, el once ...*] (The prime numbers are two, three, five, seven, eleven ...)

III- Enumerations that have expressions like "entre otras cosas", "entre otros", etc.

(21) [*Le gustan los libros, el cine, la música clásica y la pintura, entre otras cosas.*] (He likes books, movies, classical music, painters, among others things.)

2.2 About the enumerator in the enumerations

In some cases, we may detect an element that "shoots" the enumeration components. This means that it allows the enumeration inclusion in the clause. E.g. nouns can enumerate adjectival and verbal phrases, subordinate clauses, etc. In the sentence (9), "Julián" is the enumerator, because all enumeration components are linked to it, so that, the enumeration can be decomposed as shown in Figure 1.

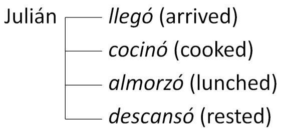

Fig. 1. Example of the enumeration decomposition.

In sentences (3) and (8) the enumerators are the verbs "bought" and "is" respectively.

In another hand, an enumeration could include a component that is the enumerator of other enumeration:

(22) [*Julián llegó, cocinó y comió pizza, hamburguesas y ensalada.*] (Julián arrived, cooked and ate pizza, hamburgers and salad.)

The last component of the first enumeration ("comió") is the enumerator of the second enumeration ("pizza, hamburguesas y ensalada").

In our work of MWE extraction, taking the enumerator in count could help us with the cases of ellipses. So, in an enumeration of bigrams conformed by the same noun in all cases, but that it is explicit only in the beginning of the series, and adjectives, it would be possible to recognize term candidates with appropriates rules. For example in an enumeration like the following:

(23) [*Derecho civil, comercial y familiar.*] (Civil, commercial and family.)

In this example really there is no adjectival enumeration, but a nominal enumeration. In this case, the enumerated elements are "derecho civil", "derecho comercial" and "derecho familiar". Nevertheless, we did not consider this kind of enumerations and we will develop rules for recognition of enumerations with enumerator in future works.

3 Related work

Research has been carried for the extraction of MWEs in monolingual [6, 19], in bilingual [7, 16, 17], and multilingual texts [3].

Piao et al. [14] use semantic taggers for MWE identification, Pecina [13] uses machine learning techniques, and Gralinski el al. [9] study the orthographic, morphological, and partially syntactic variants of contiguous MWEs. Baldwin and Kim [4] distinguish three main MWEs types. The first type is nominal MWEs, which are the most frequent in everyday language. These normally consist of a noun with its complements, such as *insuficiencia renal* (renal failure) and *crisis de asma* (asthma attack). The second type is verbal MWEs, comprising the verb and its complements, as for example *Yo he tratado la enfermedad con medicamentos* (I have treated the disease with medication). Finally, the last type is prepositional MWEs, such as *libre de culpa* (free of guilt).

In this paper, we will focus on the extraction of the nominal MWEs composed of two words (or "bigrams"), such as *insuficiencia renal* (renal failure), *ácido benzoico* (benzoic acid) and *derecho penal* (criminal law).

About enumerations, we may cite mainly the following works.

Popescu et al. [15] present enumeration rules for subclass extraction for English texts. They use this example "Biologists, physicists and chemists have convened at this inter-disciplinary conference." to argue that such rules identify "chemists" as a possible sibling of "biologists" and "physicists". They utilize two methods (SE_{self} and SE_{iter}) to count the enumerations and so extract subclass. SE_{self} method converts the number of different enumeration rules matched by each sentence of text and the average number of times that a sentence matches its corresponding rules into boolean features. SE_{iter} method use the confidence score assigned to the enumeration rules. This score is given by the average probability of extractions matched by that rule. The authors say that SE_{iter} method identifies a few more subclass than SE_{self}.

Nenadic and Ananiadou [12] describe a method to identify relationships among terms of biomedical domain in English. The method combines three text-based aspects: lexical, syntactic, and contextual similarities between terms. Lexical similarities consider the level of sharing of word constituents. Syntactic similarities use expressions described by the patterns of term enumerations and conjunctions. If two terms appear in the same expression, they are considered (highly) related. They did not discriminate the different relationships among terms, but only if the terms were related or not. Contextual similarities consider automatic discovery of relevant contexts shared among terms. As results, the authors observed that lexical and syntactic relationships have shown high precision and low recall, while contextual similarities have resulted in significantly higher recall with moderate precision.

Bosma and Vossa [5] successfully use enumerations, in addition to other criteria, as a help to detect relations between domain terms. These authors give an example, originated from the Wikipedia[1] page about frogs, which is "The skin secretions of some toads, such as the Colorado River toad and cane toad, contain ...". In this example, they show the hyponymy relation identified by identification of enumeration. One of the arguments of the relation is a single term ("toads") and the other is a list of terms ("Colorado River toad and cane toad").

With the works described, as we have seen, there are many works about MWEs and some works that use enumerations, but there are not works that use enumerations as the basis for MWE extraction. So, we believe this will be an unprecedented and promising new approach.

4 Extracting Multiword Expressions using Enumerations of Noun Phrases

Our objective then is to find out to what extent enumerations of noun phrases (NPEs) correspond to MWEs. Accordingly, we are going to identify first NPEs in our text corpora; and, second, we are going to test and evaluate extracted NPEs as MWEs.

4.1 Methodology for Multiword Expressions Extracting

The methodology for multiword expressions extracting using enumerations of noun phrases is presented in Figure 2.

The first phase is the **recognition of enumerations of noun phrase** that contains bigrams. In this phase, it is necessary determine the domain and the corpus. On the sequence, the *tokenization and morphological analysis* are carried out aiming at tagging words and punctuation marks. For this, we used the Smorph program [2] that is a finite-state part of speech (POS) tagger that Infosur[2] Group has adapted to Spanish.

The POS tagger tokenizes and categorizes the text, including punctuation marks, as is illustrated bellow. A sentence such as:

"*... presión arterial, frecuencia cardíaca y frecuencia respiratoria ...*" (blood pressure, heart rate and respiratory rate.)

will be broken down into a list of elements showed below. Smorph will provide the output[3] presented in Table 1.

[1] Wikipedia in English - `http://en.wikipedia.org`

[2] Infosur - `http://www.infosurrevista.com.ar`

[3] References: *EMS*: morphosyntactic tag, *nom*: noun, *GEN*: genre, *fem*: female, *NUM*: number, *sg*: singular, *cop*: copulative.

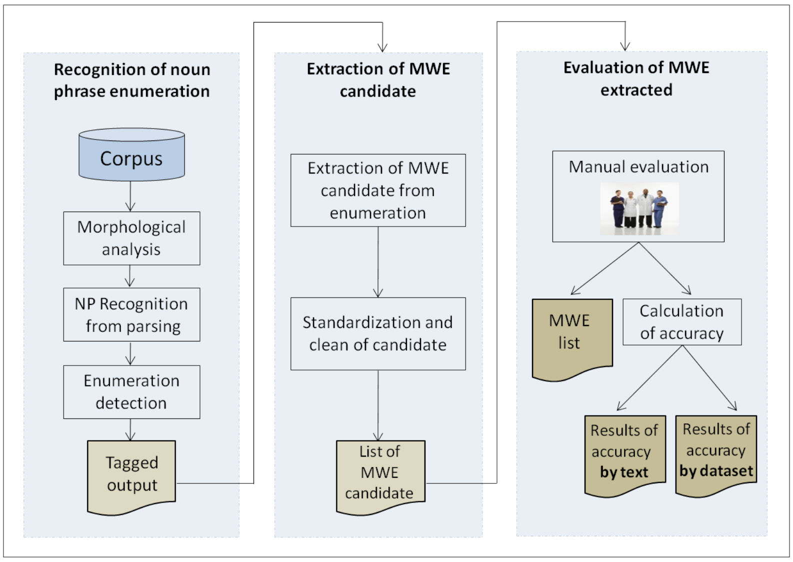
Recognition of noun phrase enumeration
Corpus
Morphological analysis
NP Recognition from parsing
Enumeration detection
Tagged output
Extraction of MWE candidate
Extraction of MWE candidate from enumeration
Standardization and clean of candidate
List of MWE candidate
Evaluation of MWE extracted
Manual evaluation
MWE list
Calculation of accuracy
Results of accuracy by text
Results of accuracy by dataset

Fig. 2. Multiword expressions extraction (MWEs) using the identification of enumerations of noun phrase.

```
'presión'.
[ 'presión', 'EMS','nom', 'GEN','fem', 'NUM','sg'].
'arterial'.
[ 'arterial', 'EMS','adj', 'GEN','_', 'NUM','sg'].
','.
[ 'cc', 'EMS','comma'].
'frecuencia'.
[ 'frecuencia', 'EMS','nom', 'GEN','fem', 'NUM','sg'].
'cardíaca'.
[ 'cardíaco', 'EMS','adj', 'GEN','fem', 'NUM','sg'].
'y'.
[ 'y', 'EMS','cop'].
'frecuencia'.
[ 'frecuencia', 'EMS','nom', 'GEN','fem', 'NUM','sg'].
'respiratoria'.
[ 'respiratorio', 'EMS','adj', 'GEN','fem', 'NUM','sg'].
```

Table 1. Output example provided by Smorph.

Once sentences have been morphologically analyzed, these become the input of a syntactic parser, which allows the identification of enumerations of noun phrase. We used the MPS syntactic parser [1] that applies a set of rules for the recognition of syntactic expressions. Firstly, we have to *detect the noun phrases* in the corpus, and we needed to create rules for that purpose. An example of NP recognition rule is “Article + Adjective + Noun = NP”, the specific MPS notation is:

S1 [L1, ‘TDET’, ‘art’] S2 [L2, ‘EMS’,‘adj’] S3 [L3, ‘EMS’,‘nom’] ⇒ S1+S2+S3 [L1+L2+L3, ‘EMS’, ‘NP’].

After this, we were able to make rules for NP *enumerations recognition*. This way; we developed a set of rules that model the enumerations of noun phrases that are described below:

– (NP + comma) ≥ 1 + NP + conjunction + NP = enumeration of noun phrase
For example: "*epidemiología clínica, anatomía patológica y diagnóstico médico.*" (clinical epidemiology, clinical pathology y medical diagnostics.).

– (NP + comma) ≥ 2 + ellipse = enumeration of noun phrase
For example: "*orden civil, defensa de los derechos, legalidad...*" (civil order, rights, legality...).

– (NP + comma) ≥ 2 + '*etcétera*' = enumeration of noun phrase
For example: "*isquemia coronaria, diabetes mellitus, asma bronquial, etcétera*" (coronary ischemia, diabetes mellitus, bronchial asthma, etcetera).

– (NP + comma) ≥ 2 + '*entre otras cosas'* = enumeration of noun phrase
For example: "*la facultad, la patria potestad, la autorización, entre otras cosas*" (faculty, parental rights, licensing, among other things).

– (NP + comma) ≥ 3 = enumeration of noun phrase (for the asyndeton cases)
For example: "*asfixia perinatal severa, parálisis cerebral, Bronquiolitis*" (severe perinatal asphyxia, cerebral palsy, Bronchiolitis).

As an example, we present the MPS notation for a possibility of the first rule:

S1 [L1, 'EMS', 'NP'] S2 [L2, 'EMS', 'comma'] S3 [L3, 'EMS', 'NP'] S4 [L4, 'EMS', 'cop'] S5 [L5, 'EMS', 'NP'] ⇒ S1+S2+S3+S4+S5 [L1+L2+L3+L4+L5, 'EMS', 'ENUM'].

With these rules, the enumerations in the corpus may be detected. Table 2 presents two examples using the sentences.

[Example 1] "*... alcanza a los textos de las leyes, decretos, reglamentos oficiales. ...*" (he reaches to the laws texts, decrees, official regulations.)

[Example 2] "*... fases evolutivas del enfisema pulmonar, bronquitis crónica y asma bronquial. ...*" (evolutionary stages of emphysema, chronic bronchitis, and bronchial asthma.)

From the *tagged output* (Table 2), which contains the detected enumerations, the **MWEs candidates are automatically extracted**. These candidates are post-processed, so they become *standardized* (in the sense of erasing numerals or any other symbol that may obstacle the recognition process). Also *stopwords* are removed at both edges of candidates.

Finally, the **evaluation of MWEs candidates extracted** is realized. Each one of these candidates is analyzed by experts of corpus domain and they indicate whether the candidate is a true MWE. This analysis is evaluated using the accuracy measurement *by text*, which is the average accuracy obtained considering the texts number of corpus, and the accuracy measurement *by corpus*, which is the accuracy total. Then, the true MWEs are stored in a list called *MWEs list*.

[Example 1]

'alcanza'. ['alcanzar', 'EMS', 'v', 'EMS', 'ind', 'PERS', '3a',
'NUM', 'sg', 'TPO', 'pres', 'TR', 'r', 'TC', 'c1', 'TDIAL',
'estrpi']. 'a'. ['a', 'EMS', 'prep']. **'los textos de las leyes ,
decretos , reglamentos oficiales'. ['el texto de el ley cc decreto
cc reglamento oficial', 'EMS', 'ENUM_NOM'**].

[Example 2]

'fases evolutivas'. ['fase evolutivo', 'EMS', 'NP']. 'del'.
['del', 'EMS', 'contr']. **'enfisema pulmonar , bronquitis
crónica y asma bronquial'. ['enfisema pulmonar cc bronquitis
crónica y asma bronquial', 'EMS', 'ENUM_NOM'].**

Table 2. Examples of enumerations detected.

4.2 Experiment

In this work, the methodology described in Section 4.1 was applied to Spanish, but it may be adapted to others languages, adjusting the linguistic informations of parsers used.

For the experiment we have narrowed the scope to bigrams (i.e. expression composed of only two words) because these are by and large the most frequent and favorable MWEs candidates. Bigrams were extracted from two different domain (see Table 3), but there is no restriction to apply the methodology in other domains. The first one is the IULA-UPF technical corpus[4] that belongs to the health and medical domains. The second one, which we will call the legal corpus, has been extracted from Wikipedia and contains texts within the taxonomy under the law category.

Corpora	**IULA-UPF**	**Legal**
Number of texts	12	20
Words average per text	8207	2558

Table 3. Description of the corpora used.

For each one of these corpus, the methodology previously showed in Figure 2 was applied. First, we realized the tokenization and morphological analysis of corpus texts using the Smorph program. Then, we used the rules developed by us as input to MPS parser. This parser utilized these rules to recognize the enumerations of noun phrases.

All bigrams contained in these enumerations are considered as candidates of MWEs. The numerals were removed from these MWEs candidates and the candidates composed only of one letter or by stopwords were discarded. The

[4] IULA-UPF technical corpus - "*Data belonging to the TECHNICAL CORPUS from Institut Universitari de Lingüística Aplicada de la Universitat Pompeu Fabra (http://bwananet.iula.upf.edu/) in December 2010.*"

stopwords from the edges of the candidates also were removed. We used the list of stopwords available in the Snowball Project[5] and we added other words, such as the full inflected paradigm of the verbs *poder* (can) and *deber* (must), adverbs as *siempre* (always), up to a total of 733 stopwords. An example of stopword remotion would be to replace the noun phrase *el médico de familia* (the family doctor) as *médico de familia* (family doctor), which is a better MWE candidate.

4.3 Results

Our success in extracting bigrams that are true MWEs has been evaluated by means of the accuracy calculation as used in the work of Gelbukh et al. [8]. In total, for the IULA-UPF corpus, we obtained 172 MWEs candidates, from which 57 candidates were wrong MWEs. Thus, the accuracy by corpus is (172-57)/172 = 67%. The true MWEs extracted are showed in Table 4. For the legal corpus, were obtained 70 MWEs candidates, out of which 28 were wrong MWEs. Thus, the accuracy by corpus is (70-42)/70 = 40%. The true MWEs extracted are showed in Table 5.

The number of words in each text (*NW.*), number of extracted MWEs candidates (*NC.*), and the accuracy by text (*P. (%)*) for each domain are presented in Tables 6 and 7. The average accuracy obtained for the IULA-UPF corpus is 69.3% and for the legal corpus is 35.4%.

5 Final Considerations

We formulated the following hypothesis: it is possible to recognize terms of a specific domain using the identification of enumerations of noun phrase. In previous experiment, we checked offers promising results for tasks of MWE extraction, and the MWEs could have the terminological unit quality for a specialized domain; such effects, the recognition of those expressions becomes very important for the automatic detection of term candidates in that domain.

Firstly, we presented an enumeration classification that was based in the components structure and in the number of the elements. From it, we took the enumerations of noun phrases and we developed a method for its recognition. In this case, we use the software SMORPH and MPS.

We tested the method with the bigrams in two different corpora. One corpus contained medical texts and the other legal texts and our results from tiny corpora are promising. With the medical corpus, the accuracy was 67% for the full corpus and 69.3% for individual texts; and with the legal corpus, the accuracy was 40% for the full corpus and only 35.4% for individual text. Poorer results with the legal corpus are explained by the fact that it contained shorter articles, and much fewer NPEs. In some cases, we could not extract bigram to test. Despite these impediments, we consider the results are encouraging, and we

[5] Stopwords list used - `http://snowball.tartarus.org/algorithms/spanish/stop.txt`

MWEs extracted from IULA-UPF corpus		
ácido benzoico	examen clínico	obstrucción nasal
ácido gástrico	expectoración mucosa	oclusiones arteriales
afecciones respiratorias	falange digital	ojeras alérgicas
alergia alimentaria	fibrosis pulmonar	parálisis cerebral
aleteo nasal	fibrosis quística	pico espiratorio
anemia falciforme	flora oral	pliegues tricipital
artritis reumatoidea	flujo espiratorio	presión arterial
asma bronquial	frecuencia cardiaca	procesos infecciosos
aspergillus fumigatus	frecuencia respiratoria	proteína catiónica
beta adrenergicos	función pulmonar	proteinosis alveolar
bloqueo simpático	función respiratoria	pruebas cutáneas
boca seca	gasimetría arterial	prurito nasal
bronquitis crónica	glándulas bronquiales	psicosis esquizofrénica
células efectoras	glándulas mamarias	pulso paradójico
células epiteliales	glutamato monosódico	reflujo gastroesofágico
colecistografía oral	heces fecales	retraso mental
complejo bradikinina	hiperreactividad bronquial	rinitis alérgica
cortisol plasmático	hipertensión arterial	rinitis crónica
cuadro clínico	historia clínica	saludo alérgico
depresión nerviosa	hongos anemófilos	secreción nasal
dermatitis atópica	iga secretoria	sexo masculino
diabetes mellitus	ige sérica	signos vitales
diagnóstico etiológico	infección viral	síndrome anginosos
dificultad respiratoria	inmunología clínica	síndrome bronquial
disfagia dolorosa	insuficiencia cardiaca	síndrome nefrótico
dolor abdominal	insuficiencia renal	sistema cardiovascular
drenaje biliar	insuficiencia respiratoria	sistema inmunitario
drogas psicoactivas	isquemia coronaria	sonidos torácicos
efectos colaterales	lengua geográfica	soplo anorgánico
enfermedad coronaria	macrófagos monocitos	tejido ectodérmico
enfermedades cardiovasculares	mecanismos inmunológicos	tensión arterial
enfermedades transmisibles	medicamentos profilácticos	tos ferina
enfisema pulmonar	medicina interna	trastornos psiquiátricos
eosinofilia periférica	médico general	tubo digestivo
eosinofilia sanguinea	moco nasal	tumor maligno
epidemiología clínica	músculo liso	ventrículo izquierdo
espasmo bronquial	neumonía bronconeumonía	verruga plana
espasmo glótico	nódulos linfáticos	vías aéreas
estudio bacteriológico		

Table 4. MWEs extracted from IULA-UPF corpus.

MWEs extracted from legal corpus		
acción reivindicatoria	derechos morales	normas jurídicas
acciones posesorias	derechos patrimoniales	orden internacional
afecciones legítimas	igualdad humanas	ordenamiento jurídico
cargos públicos	implicaciones legales	personas jurídicas
código civil	independencia doctrinal	personas naturales
cohabitación delictuosa	independencia judicial	reclamo alimentario
decretos oficiales	independencia legislativa	reglamentos oficiales
derecho tecnológico	integridad física	relaciones jurídicas
derechos conexos	negocios jurídicos	vacíos normativos
derechos humanos		

Table 5. MWEs extracted from legal corpus.

Texts	NW.	NC.	P. (%)	Texts	NW.	NC.	P. (%)
1	8531	21	81	**7**	6738	8	25
2	8604	23	87	**8**	5306	16	81
3	7820	19	79	**9**	4988	26	65
4	8377	14	71	**10**	4540	19	84
5	6926	23	48	**11**	4360	18	56
6	6604	19	79	**12**	4785	17	76

Table 6. IULA-UPF corpus results.

Texts	NW.	NC.	P. (%)	Texts	NW.	NC.	P. (%)
1	5014	1	100	**11**	677	1	100
2	16057	9	56	**12**	165	0	0
3	208	14	36	**13**	688	3	100
4	1399	5	0	**14**	11329	11	45
5	1023	3	0	**15**	448	2	0
6	189	0	0	**16**	6161	3	0
7	739	1	100	**17**	2085	7	29
8	1129	4	75	**18**	1467	1	0
9	338	3	67	**19**	815	2	0
10	141	0	0	**20**	7095	0	0

Table 7. Legal corpus results.

think that we can get better results with more recognition rules and bigger corpora.

The future work is organized according to three points. The first one is to continue working to create new rules detection for enumerations elements and, therefore, extract new candidates. The second one is to develop rules for recognition of enumerations with enumerator. The third one is to combine the extraction of enumerations of noun phrase elements with other extraction methods.

Acknowledgments

Thanks to Erasmus Mundus, CNPq, FAPESP y CONICET for financial support and to Vivaldi y Rodríguez for making available the dataset and the references lists used.

References

1. F. Abbaci. Développment du module post-smorph. In *Memória del DEA de Linguistique et Informatique*. Groupe de Recherche dans les Industries de la Langue - Universidad Blaise-Pascal - Clermont-Ferrand, 1999.
2. S. Aït-Mokhtar. *L'analyse présintaxique en une seule étape*. PhD thesis, Groupe de Recherche dans les Industries de la Langue - Universidad Blaise-Pascal - Clermont-Ferrand, 1998.
3. M. Attia, A. Toral, L. Tounsi, P. Pecina, and J. van Genabith. Automatic extraction of arabic multiword expressions. In *Proceedings of the Workshop on Multiword Expressions: from Theory to Applications*, MWE'10, pages 18–26, Beijing, China, 2010. Association for Computational Linguistics.
4. T. Baldwin and S. N. Kim. Multiword expressions. In N. Indurkhya and F. J. Damerau, editors, *Handbook of Natural Language Processing, Second Edition*. CRC Press, Taylor and Francis Group, Boca Raton, FL, 2010. ISBN 978-1420085921.
5. W. E. Bosma and P. Vossen. Bootstrapping language neutral term extraction. In *Proceedings of the 7th international conference on Language Resources and Evaluation*, LREC'10, 2010.
6. D. Català and J. Baptista. Spanish adverbial frozen expressions. In *Proceedings of the Workshop on a Broader Perspective on Multiword Expressions*, MWE'07, pages 33–40, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics.
7. J. Duan, M. Zhang, L. Tong, and F. Guo. A hybrid approach to improve bilingual multiword expression extraction. In *Proceedings of the 13th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, PAKDD'09, pages 541–547, Berlin, Heidelberg, 2009. Springer-Verlag.
8. A. Gelbukh, G. Sidorov, E. Lavin-Villa, and L. C. Hernandez. Automatic term extraction using log-likelihood based comparison with general reference corpus. In C. J. Hopfe, Y. Rezgui, E. Metais, A. Preece, and H. Li, editors, *Proceedings of the Natural Language Processing and Information Systems, and 15th International Conference on Applications of Natural Language to Information Systems*, NLDB'10, pages 248–255, Berlin, Heidelberg, 2010. Springer-Verlag.
9. F. Gralinski, A. Savary, M. Czerepowicka, and F. Makowiecki. Computational lexicography of multi-word units. how efficient can it be? In *Proceedings of the Workshop on Multiword Expressions: from Theory to Applications*, pages 2–10, Beijing, China, 2010. Coling 2010 Organizing Committee.
10. R. Jackendoff. *The architecture of the language faculty*. Linguistic inquiry monographs. Cambridge, MIT Press, 1997.
11. W. Koza. Análisis automático de textos: reconocimiento de enumeraciones (electronic version). In S. A. de Lingüística, editor, *Actas del XI Congreso de la Sociedad Argentina de Lingüística*, pages 1–11, Universidad Nacional del Litoral - Santa Fe, Argentina, 2008. Héctor Manni.
12. G. Nenadic and S. Ananiadou. Mining semantically related terms from biomedical literature. *ACM Transactions on Asian Language Information Processing*, 5(1):22–43, 2006.

13. P. Pecina. A machine learning approach to multiword expression extraction. In *Proceedings of the Language Resources and Evaluation Conference - Workshop Towards a Shared Task for Multiword Expressions*, LREC'08 - MWE'08, pages 54–57, 2008.
14. S. S. L. Piao, P. Rayson, D. Archer, A. Wilson, and T. McEnery. Extracting multiword expressions with a semantic tagger. In *Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment - Volume 18*, MWE'03, pages 49–56, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.
15. A. Popescu, A. Yates, and O. Etzioni. Class Extraction from the World Wide Web. In *Proceedings of AAAI 2004 Workshop on Adaptive Text Extraction and Mining*, ATEM'04, 2004.
16. C. Ramisch, H. de Medeiros Caseli, A. Villavicencio, A. Machado, and M. Finatto. A hybrid approach for multiword expression identification. In T. Pardo, A. Branco, A. Klautau, R. Vieira, and V. de Lima, editors, *Computational Processing of the Portuguese Language*, volume 6001 of *Lecture Notes in Computer Science*, pages 65–74. Springer Berlin / Heidelberg, 2010.
17. Z. Ren, Y. Lü, J. Cao, Q. Liu, and Y. Huang. Improving statistical machine translation using domain bilingual multiword expressions. In *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications*, MWE'09, pages 47–54, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.
18. I. A. Sag, T. Baldwin, F. Bond, A. Copestake, and D. Flickinger. Multiword expressions: a pain in the neck for nlp. In *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing'02, pages 1–15. Springer-Verlag, 2002.
19. M. Weller and F. Fritzinger. A hybrid approach for the identification of multiword expressions. In *Proceedings of the Third Swedish Language Technology Conference - Workshop on Compounds and Multiword Expressions*, SLTC'10 - MWE'10, pages 1–2, Linköping, Sverige, 2010.