

# Translation of Verbal Expressions and Context of Use Extraction Through a Corpus on Web

Arturo Velasco, María J. Somodevilla, and Ivo. H. Pineda

Facultad de Ciencias de la Computación, Universidad Autónoma de Puebla  
arturoezak@hotmail.com, {mariasg, ipineda}@cs.buap.mx

**Abstract.** The group of fixed expressions constitutes an important part of the lexical system. This group is defined as the stable combination of two or more elements that is not possible to establish a meaning from its constituents, in addition to a grammatical structure which can move away from the language rules. In this paper a method of processing, translation and context of use extraction of verbal expressions (subsets of the fixed expressions) into a diatopic system of the Spanish language is presented. The architecture proposed is organized in three modules: the database, whose objective is to be able to store essential characteristics of them; the corpus, that contains digital texts and transcribed oral language; and the extraction expressions module, which extract examples on the corpus.

## 1. Introduction

The *UFs* or *fixed expressions* are expressions consisting of two or more words whose meaning can not be inferred from the union of the significance of each of the lexical elements that constitute it. Zuluaga A., describes two basic characteristics that have fixed expressions: idiomaticity, characteristics that are peculiar and unique to a specific language or sub-language, including some socio-cultural traits; and fixation, the property that has the expressions of being reproduced in the speech like previously defined combinations, i.e. they present certain order in their syntactic structure [1].

Other important characteristics that have the *UFs* are: high frequency of report of their constituent elements, absence of grammatical rules in the expressions and translation problems.

On the other hand, due to the lack of agreement between linguists to establish limits of research of the phraseology and terminology used in this area, we decided to follow Alberto Zuluaga's work [1].

Zuluaga carries out a classification of the *UFs* based on the actions of the expressions in the speech. In the first group, Zuluaga establishes the locutions like a stable combination of two or more terms that work as an element in sentences to level of lexeme or syntagm. Inside this classification (locutions) he separates those that are in use as grammatical instruments and the expressions that possess semantic sense

(lexical units). The subset of the *UFs* object of study in this work are the *locutions* and *verbal syntagms* whom belong to the units with lexical sense. The verbal locution is equivalent to lexemes, e.g.: *pasar a mejor vida* (to die) or  *echar una mano* (to help) and the verbal syntagm are equivalent to syntagms e.g.: *pagar los platos rotos* (to suffer the consequences of something).

Considering those problems mentioned above and the *UFs* taxonomy of Zuluaga, it's proposed to develop a *Diatopic Verbal Expressions Digital Dictionary* (DIVEDD) for Spanish Language (diatopic subsystems of Spain and México) in order to enable the process of translation of verbal expressions (verbal locutions and verbal syntagms) in both subsystems. This prototype uses regular expressions and keywords, generating synonyms and variants expressions, finally, shows through a Corpus, examples of real use.

The paper is organized as follows. The second section describes related work with processing and translation of *UFs*; third section presents the architecture of the DIVEDD; the results are showed in the fourth section; and finally, conclusions are presented in the fifth section.

## 2. Related Work

The group of *UFs* constitutes an important part of the lexical system, where monolingual and bilingual dictionaries only capture certain number of units, often reduced, to an alphabetical process of selection and random description [2]. In México there are not recent works of compilation of expressions, some of them are: the *Diccionario breve de mexicanismos* [3], the *Diccionario ejemplificado de mexicanismos* [4], and the *Diccionario del español usual en México* [5]. The lack of strict rules at the time of integrate these dictionaries brought the introduction of different subsets of *UFs*.

There are some works related with translation of expressions in the Spanish language such as *Recopilación de proverbios*, proverbs which were translated into four languages (English, French, German and Italian) [6]. In *Spagnolo-Italiano: Espressioni idiomatiche e proverbi*, there are a summary of idiomatic expressions, proverbs and Spanish and Italian pragmatic sentences [7]. In [8] there are 877 *refranes españoles*, sayings with their correspondence Catalan, Galician, Basque, English and French. Finally, *Divergencias en la traducción de expresiones idiomáticas y refranes* by Corpas Pastor [9], that provides a more systematic methodology for the translation of expressions between French and Spanish (Spain). This model of bibliographical record considers different uses, the level of the speaker's registration, antonyms, synonyms, source of the expression and examples among other data. This work was considered as a starting point and taking the benefits of a corpus for showing use of actual situations of expressions.

### 3. Prototype Architecture

The architecture proposed of the DIVEDD is organized in three modules: the database, that contains the essential characteristics of the expression; the corpus, that contains in this first stage of digital texts and transcribed oral language; and the expansion expressions module, that is complemented with a list of stop-words and a database storing verbal conjugations. In the fig. 1 the architecture of the DIVEDD is shown.

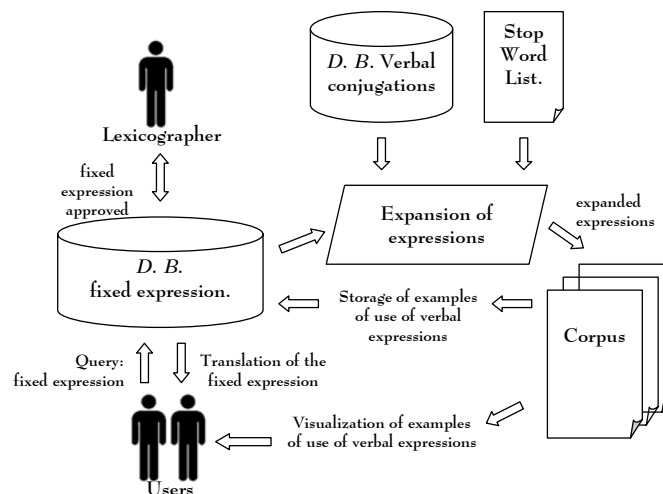


Fig. 1. DIVEDD architecture.

#### 3.1 Variants, Synonyms and +Frequents Expressions

*Variants*: those expressions that vary or omit any of its closed lexical elements without having semantic change.

*Synonyms*: those expressions that have changed in their non closed lexical element i.e. key-word or those which do not contain any element in common, but they do not have a semantic change.

*+Frequent*: the most used or most likely expression to appear in the dictionary, within a set of variations. Thus, +Frequent expression is taken as representative.

Table 1 shows an example of synonym expressions, therefore, the expressions: *ir al bote*, *ir al tambo*, *ir a la sombra* and *ir tras las rajas* are synonyms, because their keywords (*kw*) changes but they have the same definition. The same situation applies to the expressions *hacer la barba*, *hacer la pelota* and *hacer la rosca*, but in addition, *hacer la barba* is the translation into Spanish (México) of the expressions of Spain *hacer la pelota* and *hacer la rosca*.

**Table 1.** Handling synonym expressions in the DIVEDD.

	+Frequent_ Verbal_ Expression	Definition	Key-word	Thematic_ Field	Linguistic_ Record	Country
S y n o n y m s	Ir al bote	Meter a alguien en la cárcel	Bote	Behavior	Informal	México
	Ir al tambo	Meter a alguien en la cárcel	Tambo	Behavior	Informal	México
	Ir a la sombra	Meter a alguien en la cárcel	Sombra	Behavior	Informal	México
	Ir tras las rejas	Meter a alguien en la cárcel	Rejas	Behavior	Informal	México
	Hacer la barba	Lisonjear a alguien	Barba	Behavior	Informal	México
	Hacer la pelota	Lisonjear a alguien	Pelota	Behavior	Informal	Spain
	Hacer la rosca	Lisonjear a alguien	Rosca	Behavior	Informal	Spain

Table 2 shows variants through regular expressions. A regular expression is a set of pattern matching rules encoded in a string according to certain syntax rules [10]. Thus, it is possible to describe or represent a set of strings without need to enumerate all of its elements. The operators used in the right column of Table 2 are described in table 4 in section 3.4, *Generation of Synonyms and Variants*.

**Table 2.** Variants expressions in the database of the DIVEDD.

<i>+Frequent_ Verbal_ Expression</i>	<i>Definition</i>	<i>Key-word</i>	<i>Variants_Verbal_Expression</i>
Ir al bote	Meter a alguien en la cárcel	Bote	[ir,llevar,meter] (al) {bote} [refundir] (en_el) {bote}
Ir al tambo	Meter a alguien en la cárcel	Tambo	[ir,llevar,meter] (al) {tambo} [refundir] (en_el) {tambo}
Ir a la sombra	Meter a alguien en la cárcel	Sombra	[ir,llevar,meter] (a_la) {sombra} [refundir] (en_la) {sombra}
Ir tras las rejas	Meter a alguien en la cárcel	Rejas	[ir,llevar,meter,refundir] (tras_las) {rejas}

### 3.2 The Database

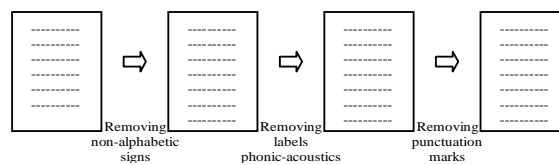
This module is based on a relational model that provides mechanisms that guarantee to avoid duplicity of records and inconsistency problems; it also guarantees the referential integrity and favors improvements of processing of the expressions. Table 3 shows the most important attributes of verbal expressions to store.

**Table 3.** More important attributes of the verbal expressions.

ATTRIBUTE	DESCRIPTION
<i>Verb</i>	Main verb used in the expression.
<i>Canonical_ Verbal_ Expression</i>	<i>Locution</i> or <i>verbal syntagm</i> in its canonical form.
<i>Definition</i>	Definition of the verbal expression. Field used to make the translation among the diatopic verbal expressions.
<i>Source</i>	Resource where the expression was extracted.
<i>Use_ Frequency</i>	The number of frequencies of appearance of the expression in the corpus.
<i>Linguistic_ Record</i>	Level of registration of the expression.
<i>Country</i>	Country of origin of the expression.
<i>Region</i>	area or region of use of the expression.
<i>Thematic_ Field</i>	Thematic field of the expression
<i>Key-Word</i>	Alexical component unit of the expression. Useful to distinguish among synonym expressions.
<i>Variant_ Verbal_ Expression</i>	Field of the table <i>Variant</i> that stores the variants of the canonical verbal expressions.
<i>Example</i>	Field of the table <i>Examples</i> that stores the examples provided by the lexicographer and extracted of the corpus.

### 3.3 Corpus Processing for DIVEDD

The corpus of the DIVEDD is conformed by written language and transcribed oral language [11], based on the recommendations of Sinclair J. [12], [13]. The subcorpus of written language is built from digital Mexican newspapers (four sections: *local news*, *police section*, *opinion section* and *shows section* over geographical limitation in the states of D.F., Mexico, Hidalgo, Morelos, Puebla and Tlaxcala). The subcorpus of transcribed oral language is conformed by the Sociolinguistic Corpus of the México City (CSCM) [14]. Fig. 2 shows the processing for the last one subcorpus.



**Fig. 2.** Corpus processing.

### 3.4 Extraction Expressions Module

The module expressions expansion serves as a liaison between the DB and the corpus of DIDEVD. This module has two contributions: generate synonyms and variations of the +Frequent expression stored in the DB; and extract examples of actual usage throw the corpus.

**Generation of Synonyms and Variants.** The generation of synonyms and variants of a *+frequent* expression requires the entry of the possible combinations that can occur between *kw*'s and *connectors-words* or *stop-words* (*sw*). To describe all these expressions without the need to enumerate each one of them, the regular expressions are used. The operators are shown in Table 4.

**Table 4.** Operators of the regular expressions.

Operator	Function
'[ ' y ' ]'	Denotes the set of verbs that can be used in the expression.
'( ' y ' )'	Denotes the set of connector-words between the verb and key-word.
'{ ' y ' }'	Denotes the set of key-words that are used in the expression.
','	Separator of a set of words (verbs, key-words, connector-words). Can be use ',' instead of ' '.
'	Performs the same function as ','.
_'	Joint two or more nonseparable words in an expression
' '	The blank space denotes the separation between groups.

Considering the operators used in regular expressions specified in Table 4, the canonical verbal expression formed by *hablar más que un loro*, where the *kw* is *loro* and the regular expressions are denoted by:

[hablar,platicar] (más\_que\_un,como,como\_un) {loro,perico,merolico}  
 [hablar,platicar] (más\_que\_una,como\_una) {cotorra}

The set of variants of *hablar más que un loro* are: *hablar como loro*, *hablar como un loro*, *platicar más que un loro*, *platicar como loro* and *platicar como un loro*.

The set of all its synonyms are: *hablar más que un perico*, *hablar como perico*, *hablar como un perico*, *platicar más que un merolico*, *platicar como merolico*, *platicar como un merolico*, *hablar más que una cotorra*, *hablar como una cotorra*, *platicar más que una cotorra* and *platicar como una cotorra*.

As it can be seen the properties of regular expressions help us to match any possible variation of the expressions in the corpus without necessity of having enumerated each one.

**Extraction of Usage Examples.** The second contribution of extraction expressions module is the search for examples of real use of expressions stored in the DB.

The search can be performed by matching between the expression and a fragment of the corpus. The second way is to find expression through similarity. Based on the premise that the *kw* is crucial in the expression,

The first step is to identify all the words that have a high degree of similarity with the *kw* was carried out. The similarity function between two strings is described in [15] and showed below.

The second step is to identify words that preceding to the *kw*. Only *verbs* and *sw*'s are accepted. Another word unidentified will provoke that the fragment is rejected. The process consists of a retreat from the position of the *kw*. Ends in a satisfactory manner when encountering a *verb* and *sw*'s or if two verbs (an auxiliary and non auxiliary) and *sw*'s are matched.

```
//Similarity of two strings; return the percentage
function similarity($s1, $s2)
{
    $m = strlen($s1);
    $n = strlen($s2);
    $matrix = array(array($m),array($n));
    for($i=1; $i < $m; $i++) $matrix[$i][0]=0;
    for($j=0; $j < $n; $j++) $matrix[0][$j]=0;
    for ($i=1; $i <= $m; $i++) {
        for ($j=1; $j <= $n; $j++) {
            if ($s1[$i-1]==$s2[$j-1])
                $matrix[$i][$j] = $matrix[$i-1][$j-1] + 1;
            else if ($matrix[$i-1][$j] >= $matrix[$i][$j-1])
                $matrix[$i][$j] = $matrix[$i-1][$j];
            else
                $matrix[$i][$j] = $matrix[$i][$j-1];
        }
    }
    $avgs = ($m + $n) / 2;
    return ($matrix[$m][$n] / $avgs) * 100;
}
```

#### 4. Results

The extraction process using regular expressions have shown little recovery since it requires a tie with the exact expressions given by the lexicographer. On another hand, the extraction process by similarity functions between *kw*'s and words in the corpus heralds not only an examples extraction process; also, variant expressions can be extracted to perform in a future work, the reverse process, i.e. creating regular expressions.

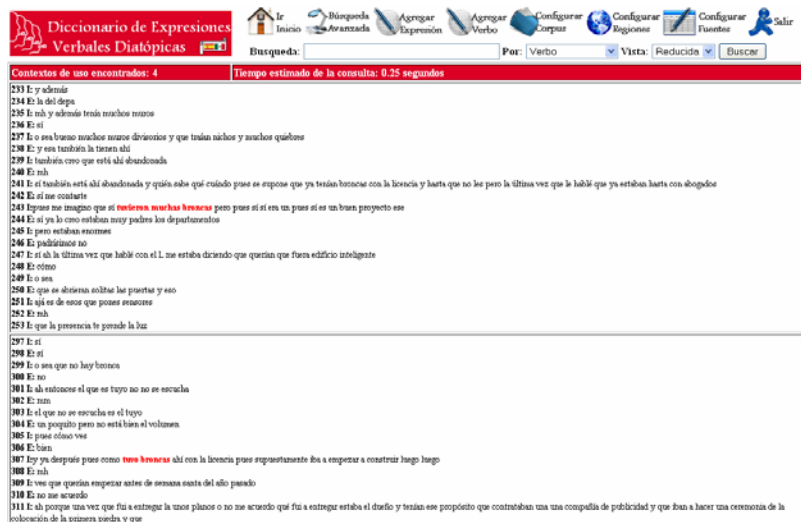
Thus, besides of the similarity applied and described in [15], the Levenshtein similarity was applied. This function calculates the minimum number of operations (insertion, deletion or substitution) required to transform one string into another. The results are not as regulars as [15], and has trouble distinguishing different *kw*'s, and grouping *kw*'s of which varies only in gender and number.

The DIVEDD database was developed in MySQL 5.0.18. All the processing is implemented in PHP 5.1.2. The fig. 3 shows the DIVEDD interface.



Verbo	Expresión Verbal	Definición	P. Clave	Camp. Tem.	Niv. Reg.	País		
aventar	aventar a alguien flores	adular, lisonjear	flores	comportamiento	culto	México		
cater	cater bien a alguien	obtener buena acogida	bien	comunicación	culto	México		
chingar	chingar algo o a alguien	decomponer algo, importunar, molestar	chingar	comportamiento	informal	México		
dar	dar a alguien el avión	no prestar atención	avión	comportamiento	culto	México		
dar	dar chance	dar permiso, oportunidad	chance	comunicación	informal	México		
dar	dar la cara	salir a su defensa	madre	comportamiento	estándar	México		
dar	dar lana a alguien	dar dinero	lana	comunicación	informal	México		
dar	dar un rol	parar, viajar	rol	comunicación	estándar	México		
dejar	dejar algo botado	demitirle importancia	botado	comportamiento	estándar	México		
dejar	dejar algo o alguien por la paz	no inquietarle ni molestarle, dejar en paz a alguien	paz	comportamiento	culto	México		
dejar	dejar plantado a alguien	hacer esperar a alguien sin acudir a la cita	plantado	comportamiento	estándar	México		
echar	echar la culpa	atribuirle la falta o delito que se presume ha cometido	culpa	comportamiento	estándar	México		
estar	estar algo cabrón	difícil, complicado	cabrón	descripción	informal	México		
estar	estar algo canijo	difícil, Mala persona	canijo	descripción	estándar	México		
estar	estar algo cañón	muy bien, estupendo	cañón	descripción	informal	México		
estar	estar algo chado	bueno, muy bueno	chado	descripción	estándar	México		
estar	estar algo chingón	bueno, muy bueno	chingón	descripción	informal	México		
estar	estar algo en chino	difícil de entender	chino	descripción	culto	España, México		
estar	estar algo muy colgado	lejano en distancia, tiempo	colgado	descripción	estándar	México		
estar	estar algo muy pesado	violento, insoportable, difícil de soportar	pesado	descripción	estándar	México		

Fig. 3. DIVEDD interface.



Diccionario de Expresiones Verbales Dialectales		Inicio	Búsqueda Avanzada	Agregar Expresión	Agregar Verbo	Configurar Corpus	Configurar Regiones	Configurar Fuentes	Salir
Contextos de uso encontrados: 4		Tiempo estimado de la consulta: 0.25 segundos							
233	E: y además								
234	E: la del deya								
235	E: mh y además tenía muchos sonos								
236	E: ei								
237	E: o sea bueno muchos sonos divertidos y que traían náchos y muchos quibnos								
238	E: y eso también la tenían ahí								
239	E: también como que está ahí abandonada								
240	E: mh								
241	E: el también está ahí abandonada y quién sabe qué cuando pasa se espone que ya tenían broncas con la licencia y hasta que no les pero la última vez que la había que ya estaban hasta con abogados								
242	E: si me cuentan								
243	E: pero me imagino que si <b>tenían broncas</b> pero si en un caso si es un buen proyecto sea								
244	E: si ya lo caso estaban muy padre los departamentos								
245	E: pero estaban sonos								
246	E: quédándose no								
247	E: si ahí la última vez que habló con el L. me estaba diciendo que quería que fuera edificio inteligente								
248	E: cómo								
249	E: o sea								
250	E: que se abrense así las pasadas y eso								
251	E: ahí es de esos que pasan sonos								
252	E: mh								
253	E: que la presencia te prende la luz								
257	E: ei								
258	E: ei								
259	E: o sea que no hay broncas								
260	E: no								
261	E: ahí entonces el que se tuyo no se escuchó								
262	E: mm								
263	E: el que no se escuchó es el tuyo								
264	E: un poquito pero no está bien el volumen								
265	E: pues cómo ves								
266	E: bien								
267	E: y después pasa como <b>tenían broncas</b> ahí con la licencia pasa supuestamente los a empezar a construir luego luego								
268	E: mh								
269	E: vea que querían empezar antes de semana santa del año pasado								
270	E: no me acuerdo								
271	E: ahí porque una vez que fui a entregar la tasa planeo o no me acuerdo qué fui a entregar estaba el diario y tenían ese propósito que contrataban una una compañía de publicidad y que iba a hacer una ceremonia de la colocación de la primera piedra y que								

Fig. 4. Extraction of real examples of *tener broncas*.

## 5. Conclusions and ongoing work

The DIVEDD appears to be a system for human translation assisted by computer, providing a definition and basic characteristics of verbal expressions. The DIVEDD does not try to be a detailed dictionary but it is as a mechanism reliable of storage of



phraseological information enforcing structure and integrity of data, reducing times of search and translation. On the other side, the mechanisms of search expressions by expressions' attributes and its combinations make the DIVEDD a flexible tool.

Finally, note that the extraction process starting from similarity on *kw's* showed more encouraging results, but with *noise* (information not relevant to the query), because there are parts of texts recovered, that are not relevance to the phraseology, i.e. there are non-verbal expressions. We will work on linguistic heuristics to reduce the noise.

## References

1. Zuluaga A. Introducción al estudio de las expresiones fijas. Frankfurt: Peter Lang. (1980)
2. Mogorrón H. Los diccionarios electrónicos fraseológicos, perspectivas para la lengua y la traducción. Universidad de Alicante (2004)
3. Gómez de Silva G. Diccionario breve de Mexicanismos. 1a ed., México, FCE. (2001)
4. Steel B. Breve Diccionario Ejemplificado de Mexicanismos (2000)
5. Lara L. Diccionario del español usual en México. 1a ed. ISBN: 9789681207045 (2003)
6. Casado, M. L., Agueda, S., Agueda, B. and Corral, J. Recopilación de proverbios. Alcobendas. ISBN: 9788471436450 (1998)
7. Zamora M. Spagnolo-italiano: espressioni idiomatiche e proverbi, Milano, EGEA (1997)
8. Sevilla M. and Cantera J. 877 refranes españoles con su correspondencia catalana, gallega, vasca, francesa e inglesa. Madrid: EUNSA (1998)
9. Sevilla M. Divergencias en la traducción de expresiones idiomáticas y refranes (francés-español) (1999)
10. Stubblebine T. Regular Expression, Pocket referente. O'really 2nd edition (2007)
11. Procházková P. Fundamentos de la lingüística de corpus, concepción de los corpus y métodos de investigación con corpus (2006).
12. Sinclair J. Preliminary Recommendations on Corpus Typology. EAGLES Document EAG-TCWG-CTYP/P. (1996)
13. Sinclair J. Developing Linguistic Corpora: a Guide to Good Practice Corpus and Text—Basic Principles (2004)
14. Colegio de México, <http://lef.colmex.mx/Sociolingüística/CSCM/Corpus.htm>
15. Oliver J. DecisionGraphs-An Extension of Decision Tres. TechnicalReport No:92 /173 (1993)