Recibido / Received: 30/10/2019
Aceptado / Accepted: 28/02/2020

Para enlazar con este artículo / To link to this article:
http://dx.doi.org/10.6035/MonTI.2020.ne6.3

Para citar este artículo / To cite this article:
KOZA, Walter & Natalia RIVAS FOLCH. (2020) "Computational modelization of verbal idioms of Chilean Spanish based on a lexicon-grammar proposal." In: MOGORRÓN HUERTA, Pedro (ed.) 2020. *Análisis multidisciplinar del fenómeno de la variación fraseológica en traducción e interpretación / Multidisciplinary Analysis of the Phenomenon of Phraseological Variation in Translation and Interpreting*. *MonTI* Special Issue 6trans, pp. 94-118.

COMPUTATIONAL MODELIZATION OF VERBAL IDIOMS OF CHILEAN SPANISH BASED ON A LEXICON-GRAMMAR PROPOSAL

WALTER KOZA
walter.koza@pucv.cl
Pontificia Universidad Católica de Valparaíso
Proyecto FONDECyT 1171033

NATALIA RIVAS FOLCH
natalia.rivas.f@gmail.com
Pontificia Universidad Católica de Valparaíso
Proyecto FONDECyT 1171033

Abstract

This work is a formal study of Verbal Idioms (VI) in Chilean Spanish, proposed with the objective of developing a computational algorithm for the automatic analysis of texts. For this, the VIs listed in the *Diccionario de Uso del Español de Chile* were described according to lexicon-grammar theory. Computational modeling was then performed with NooJ, a program that has several utilities. This method was evaluated for automatic recognition in a corpus composed of texts from Chilean press sources. The model obtained 96.4% precision, 92.23% coverage, and 94.53% in F-measure. The results show that the algorithm designed is adequate for automatic analysis of VIs.

Keywords: Verbal idioms. Chilean Spanish. Automatic Analysis. Lexicon-grammar. NooJ.

Resumen

En el presente trabajo, se propone un estudio formal de las locuciones verbales (LV) pertenecientes al español de Chile con el propósito de diseñar un algoritmo computacional para el análisis automático. Para ello, se describen las VI incluidas en el *Diccionario de Uso del Español de Chile*, según los criterios de la léxico-gramática. Posteriormente, se procede a la modelización computacional recurriendo a NooJ, un programa que cuenta con diversas utilidades para el análisis lingüístico. Este método fue evaluado mediante el reconocimiento automático en un corpus compuesto por textos de prensa chilena. Se obtuvo 96,4% de precisión, 92,23% de cobertura y 94,53% de medida F. Los resultados demuestran que el algoritmo diseñado es adecuado para el análisis automático de las VI.

Palabras clave: Locuciones verbales. Español chileno. Análisis automático. Léxico-gramática. NooJ.

1. Introduction

One of the most relevant problems that computational linguistics faces is to establish an effective way to analyze speech (Vietri 2014, Silberztein 2016). The complexities involved in automatic analysis of these expressions are fundamentally related to the degree of fixation, since while some expressions do not allow their elements to be separated, others do, as in (1), below. Additionally, another inconvenience derives from cases with modifiable elements (2). And in the case of verbal idioms, a challenge lies in the fact that the verb can acquire syntactic conditions that it does not possess when it appears freely (3):

(1) Juan **agarra para el hueveo** a María. → Juan **agarra** a María **para el hueveo**.
(2) Juan **dejó la escoba** → Juan **dejó la *mansa* escoba**.
(3) Juan tiene una casa / *La casa es tenida por Juan. → **Juan tiene en cuenta a María**. / María **es tenida en cuenta** por Juan.

To this end, this paper proposes to analyze verbal idioms (VIs) in Chilean Spanish with a view to designing an automatic analysis computational algorithm for natural language texts. This objective requires an exhaustive description of the morphosyntactic behavior of these elements,

under the theoretical and methodological approaches of lexicon-grammar theory (Gross 1975; 1984). Our data set is made up by the VIs listed in the Dictionary of Spanish Usage in Chile (*Diccionario de uso del español de Chile*, DUECh) (Academia Chilena de la Lengua 2010) to analyze and categorize them according to their structure and in relation to the arguments and transformational possibilities they have.

The process described above is computer-modeled using the open-access software NooJ, developed by Silberztein (2003; 2016). This program presents a structure based on the Chomsky-Schützenberger hierarchy (Chomsky 1957) and has several utilities for linguistic analysis. The computational work involved the construction of an electronic dictionary and syntactic grammars of VI. This method was evaluated by means of automatic recognition in a corpus composed by texts from a Chilean press source, achieving 96.4% precision, 92.23% coverage, and 94.53% F-measure. These results demonstrate that the designed algorithm is suitable for the automatic analysis of the VIs.

2. Theoretical Framework

This section presents the theoretical framework that supports the research. It includes, on the one hand, works referring to the nature of each idiom and, on the other hand, the proposals of lexicon-grammar theory (Gross 1975; 1984). This latter includes the work of Vietri (2014), who carries out an analysis of Italian VIs and, as in this study, designed a computational algorithm in NooJ.

2.1. Regarding the concept of idiom

In addition to free word combinations, it is possible to find in language systems certain fixed structures that possess greater stability. These are structures that, according to Corpas Pastor (1996: 18):

i. Consist of at least two spelling words.
ii. Present a certain degree of lexicalization.
iii. Are characterized by a high frequency of coappearance in that language.

These constructions are called *phraseological units* (PU) by the author and have enjoyed great attention in linguistics. Within these expressions we include idioms, understood - according to the proposal of Casares (1950) and reformulated by Penadés (2012) - as "the fixed combination of words that functions as a sentence element and whose meaning does not correspond to the sum of the meanings of its components" (Penadés 2012: 25).

The most commonly used criterion for phrase classification refers to its grammatical category, identified according to its syntactic function, i.e., nominal, adjectival, prepositional, adverbial, and verbal (Zuluaga 1975; Corpas Pastor 1996; Penadés 2012). Verbal idioms contain a verb and form a complex predicate when taking a complement or adjunct. This implies that VIs: (i) require arguments; (ii) are subject to the restrictions of concordance by means of inflection; (iii) can designate states and events (Silvagni 2017); and (iv) may be modified by an adjunct.

However, in relation to item (i), in order to identify the arguments required by VIs, it is necessary to take into account the restrictions that the idiom imposes on its arguments. Thus, for instance, '*calentar*' in '*calentar el asiento*' only requires a single argument that must have the +human feature, since it presents a crystallized direct complement:

(4) Juan calentó el asiento.
*El sol calentó el asiento.

Likewise, it is also pertinent to observe transformational possibilities in VIs, such as passive voice, nominalization, etc.

(5) Juan agarra a María para el hueveo = María fue agarrada para el hueveo por Juan.
Pedro agachó el moño = Pedro tuvo una agachada de moño.

For a comprehensive description of both the nature of arguments and the transformational possibilities in VIs, we used the lexicon-grammar framework, the main postulates of which are presented in the following section.

2.2. Lexicon-grammar theory

Lexicon-grammar theory (Gross 1975; 1984) is a method for describing natural languages based on three main aspects: (i) the syntax is indissoluble

from the lexicon; (ii) the minimum unit of analysis is the simple sentence (as opposed to words or phrases); and (iii) the framework must offer a formalization and a descriptive method applicable to any language (Elia, Monteleone & Marano 2011). This model proposes a combined study of the syntactic rules and preferences of lexical selection in relation to the transformation possibilities of a given sentence.

In order to arrive at such a description, Gross (1975) proposes the elaboration of tables organized into classes of lexical elements of shared grammatical properties (Tolone 2012). A table is presented in the form of a matrix, which contains: (i) per row, the elements of the corresponding class; (ii) per column, the syntactic-semantic properties (which are not necessarily accepted by all the members of the class); and (iii) at each intersection, a + or -, depending on whether or not the lexical entry described by the line accepts the property described by the column. A syntactic-semantic property is information that directly refers to the base construction associated with the class, to a transformation of the base construction, or to an additional construction. If a lexical unit has two different meanings, then it will present two entries. By way of example, table 1 shows a fragment of one of the French verb tables adapted by Pivaut (1989), presented in Toulouse (2012).

N0 =: Nhum	N0 =: N-hum	N0 =: Nnr	<ENT>Ppv	Ppv =: se figé	Ppv =: les figé	Ppv =: Neg	<ENT>V	Neg	N0 V	N0 être V-ant	N0 V de N0pc	N1 =: Nhum	N1 =: N-hum	N1 =: le fait Qu P	Ppv =: lui	Ppv =: y	[extrap]	NOidée V Loc N1esprit	<OPT>
+	–	–	*les*	–	+	–	*lâcher Advm*	–	+	–	–	+	–	–	–	–	–	–	*Max les lâche difficilement à Ida*
+	–	–	*<E>*	–	–	–	*renaître*	–	+	+	–	–	+	–	–	+	–	+	*Max renaît au bonheur de vivre*
+	–	–	***se***	+	–	–	***rendre***	–	+	–	–	+	**+**	+	–	+	+	–	***Max s'est rendu à mon opinion***
+	–	–	***se***	+	–	–	***rendre***	–	+	–	–	+	**-**	–	–	–	–	–	***Le caporal s'est rendu à l'ennemi***
+	–	–	*<E>*	–	–	–	*renoncer*	–	–	–	–	+	+	–	–	+	–	–	*Max renonce à son héritage*
+	+	+	*ne*	–	–	+	*revenir*	+	–	–	–	+	–	–	+	–	–	–	*La tête de Luc ne revient pas à Max*

Table 1. Example of a Lexicon-Grammar table (Tolone 2012).

As you can see, a verb such as *rendre* ('to surrender') will have double entries depending on the types of objects it takes (N0=Human, N1=Non-Human:

Max se rindió a mi opinión ('Max surrendered to my opinion'); N0=Human, N1=Human: *El cabo se rindió al enemigo* ('The corporal surrendered to the enemy').

The transformations that each sentence admits are then identified in the last column (<OPT>). The notion of transformation is proposed by Harris (1970) and alludes to the diverse possibilities in which an argument structure can be projected. This idea should not be confused with the idea of transformative generative grammar – in which the Harrisian perspective suggests that all transformations take place in the surface structure. In the example presented, it can be seen that the two sentences admit nominalization ('Max's surrender to my opinion took place', 'The corporal's surrender to the enemy took place').

Gross (1975) points out that this methodology can be extended to idiomatic expressions. On the basis of this theoretical-methodological apparatus, Vietri (2014) carried out an analysis of Italian VIs, starting with introspection and linguistic judgments, then corroborating these two criteria with information from the Internet. Following the description of the units, the author explained the possibilities introduced by lexicon-grammar theory in order to formalize the idiomatic expressions, then made use of software NooJ to create grammars and detect the properties of the units and the respective variations admitted by them.

In this sense, some of the conclusions reached by Vietri (2014) are (i) VIs are governed by the same syntactic rules as predicative verbs; (ii) idioms may be deconstructed, but only on a case-by-case basis; (iii) the semantic representation of idiomatic expressions is directly linked to their syntax; and (iv) the use of lexicon-grammar theory constitutes an adequate framework for the exhaustive description of units such as those studied, especially as concerns subsequent formalization of rules for computational modeling.

Here we take as a basis the proposal of this author, and apply it to VIs in Chilean Spanish. Following that study, we make a morphosyntactic description of these units, including their combinatory possibilities through the argument structure they project. Likewise, their transformational possibilities are considered (nominalization, passive voice, etc.). Subsequently, we

propose a formalization and computational modeling for automatic detection in natural language texts.

3. Methodology

This section presents the methodological guidelines that support this research. It is composed of two sub-sections: the first describing the nature of VIs; and the second, computational work.

3.1. Description of VIs

Descriptions were compiled from a list of VIs extracted from the DUECh (Chilean Academy of the Language 2010). This dictionary concentrates a series of lexical entries or lemmas that refer to expressions used in the speech of Chile. According to Matus, director of the dictionary, the work aims to reflect the current, socially stabilized use of the language of the country (Sáez 2012). In this sense, we considered VI every expression indicated as such in the DUECh, beyond certain omissions indicated by Sáez (2012). However, this work aims to provide a method to morphosyntactically describe any fixed expression made up of a verb plus arguments or fixed adjuncts, to be applied to new lexical units of this type detected or created in the future.

The first step was the extraction of all entries under the category "verbal idiom", for a total of 833 expressions collected. They were then classified into two large groups: VIs that expressed no denial ('*agachar el moño*') into Group 1; and those that did ('no ver una') into Group 2.

	Total lemmas	**Percentage**
Idioms Group 1	792	95.1 %
Idioms Group 2	41	4.9 %
Total idioms (G1 + G2)	833	100 %

Table 2. Summary of idioms contained in DUECh.

Table 2 shows a summary of the total number of phrases extracted from the DUECh. The present work describes the first group, leaving the second for future research.

After compilation, structures were categorized and described according to the criteria of lexicon-grammar theory (Gross 1975; 1984). This procedure is detailed below.

3.1.1. Distributional analysis

The first instance of analysis took the way in which the lemma is structured in the dictionary into consideration, that is, observing the constituent elements of the idiom. In this way, expressions such as *dejar la escoba* and *sacarse el pillo* were grouped into different classes, since one requires a clitic and the other does not. However, certain idioms were shown to require different arguments than initially contemplated (Table 3). Thus, those idioms declared in the DUECh with an enclitic pronoun, such as *meterse en un queso* could be used without it in cases such as (4):

(4) Juan *metió en un queso* a María.

In the same way, in idioms whose lemma appears without the pronoun, as in *sacar en cara*, its use does require it:

(5) María te *sacó en cara* todo lo que había hecho por ti.

This analysis yielded the following classes:

Class	Example phrase	Example use	Total	%
Cl 1: (N0)+V+C1	*echar al agua.*	*María echó al agua a Juan.*	405	51.14
Cl 2: (N0)+CL+V+C1	*arrancarse con los tarros.*	*Juan se arrancó con los tarros.*	287	36.24
Cl 3: (N0)+CL+CL+V+C1	*sabérselas por libro.*	*Ella jura que se las sabe por libro.*	6	0.76
Cl 4: (a)+(N0)+CLdat+V	*crujirle a alguien.*	*A Juan le crujió por fin.*	6	0.76
Cl 5: (a)+(N0)+CLdat+V+C1	*faltarle a alguien chauchas para el peso.*	*A María le faltan chauchas para el peso.*	35	4.42
Cl 6: (a)+(N0)+se+CLdat+V+C1	*aconchársele a alguien los meados.*	*A Juan se le aconcharon los meados, por eso no reacciona.*	40	5.05
Cl 7: (a)+(N0)+se+CLdat+V	*chupársele a alguien.*	*A María se le chupa ir a hablarle.*	3	0.38
Cl 8: (a)+(N0masc)+CLdat+V+C1	*gustarle a un hombre las patitas de chancho.*	*Está claro que a Juan le gustan las patitas de chancho.*	2	0.25
Cl 9: (a)+(N0masc)+se+CLdat+V+C1	*quemársele a un hombre el arroz.*	*María no sabía que a Juan se le quemaba el arroz.*	7	0.88
Cl 10: (a)+(N0fem)+CL+V+C1	*dejar a una mujer el tren.*	*A María la dejó el tren de tanto esperar a Juan.*	1	0.13

Table 3. Summary, phrase categorization (Group 1).

Table 3 shows how various categories emerged from the various argumentative requirements of the phrases. The prototypical organization of the

expression when used is shown for each class. Likewise, each category contemplates a series of elements contained in the idiom:

Element	Meaning
(a)	Introduces the indirect complement in the idiom.
(N0)	First argument required by the idiom, which can have both a subject and an indirect complement function. An N0 can be tacit and have no phonological realization, which is indicated in parentheses.
le/se	Accounts for the clitic pronouns determined or defined by a particular idiom. In these cases, the expression only admits these clitic options.
CL	Indicates that a dative or accusative clitic pronoun can be inserted in that part of the phrase. However, some phrases require a particular type of pronoun; in those cases, it is indicated in the class as "CLdat", since only dative pronouns are allowed there.
V	Verbal form that must be conjugated. It is constituted as the core of the idiom.
C1	Represents the "invariable" part of the idiom.

Table 4. Description of the constituent elements.

After categorization, we continued with the preparation of lexicon-grammar theory tables for each class, creating a database of the different arguments for the VIs. The structures of the tables were adapted following Gross (1984) and Vietri (2014), as shown in the excerpt from the distributional analysis of class 5 VIs, below.

PREP	N0		PronCl			Verbo	CS	CS						PREP						Ejemplo
a	+Anim	-Anim	Se	Dativo	Acusativo			SD	SN	SP	SA	SAd	SV	a	de	con	por	para	N1	
(+)	(+)	-	-	+	-	bajarle	la indiada.	+	-	-	-	-	-	-	-	-	-	-	-	A Juan le bajó la indiada, después se taimó y se fue.
(+)	(+)	-	-	+	-	caerle	el tejazo.	+	-	-	-	-	-	-	-	-	-	-	-	Finalmente, a Juan le cayó el tejazo
(+)	(+)	-	-	+	-	caerle	el tejo.	+	-	-	-	-	-	-	-	-	-	-	-	Hasta que le cayó el tejo a Juan
(+)	(+)	-	-	+	-	caerle	la chaucha.	+	-	-	-	-	-	-	-	-	-	-	-	A Juan le cayó la chaucha, se demoró pero ya se dio cuenta
(+)	(+)	-	-	+	-	caerle	la teja.	+	-	-	-	-	-	-	-	-	-	-	-	A María le cayó la teja y encaró a Juan
(+)	(+)	-	-	+	-	comerle	las manos.	+	-	-	-	-	-	-	-	-	+	-	+	A María le comen las manos por volver a competir
(+)	(+)	-	-	+	-	crecerle	el pelo.	+	-	-	-	-	-	-	-	-	-	-	-	A Pedro le creció el pelo, ya no pertenece a este lugar
(+)	(+)	-	-	+	-	crujirle	el mate.	+	-	-	-	-	-	-	-	-	-	-	-	Hasta que a María le crujió el mate y se dio cuenta de cómo eres
(+)	(+)	-	-	+	-	crujirle	la mollera.	+	-	-	-	-	-	-	-	-	-	-	-	A Juan le crujió la mollera, gracias a eso se solucionó el problema

Table 5. Example, lexical grammar table for class 5 VIs.

These lexicon-grammar theory tables take the form of a matrix, with rows defining each of the idioms, and columns specifying the diverse syntactic-semantic properties. Next, the number of VIs included in each class is summarized.

Class	Total phrases	Total Entries
Cl 1: (N0)+V+C1	405	798
Cl 2: (N0)+CL+V+C2	287	459
Cl 3: (N0)+CL+CL+V+C3	6	12
Cl 4: (a)+(N0)+CLdat+V	6	6
Cl 5: (a)+(N0)+CLdat+V+C5	35	36
Cl 6: (a)+(N0)+se+CLdat+V+C6	40	40
Cl 7: (a)+(N0)+se+CLdat+V	3	4
Cl 8: (a)+(N0masc)+CLdat+V+C8	2	2
Cl 9: (a)+(N0masc)+se+CLdat+V+C9	7	7
Cl 10: (a)+(N0fem)+CL+V+C10	1	2

Table 6. Comparison of number of lemmas per class versus number of entries in the lexical-grammar table.

The classes differ greatly between the number of lemmas and dictionary entries. If classes were placed on a continuum – where 1 is the simplest, and 10, the most complex – as classes become more complex, the uses or properties they allow are more restricted. This translates into a reduction in the number of entries in the tables.

3.1.2. Transformations

Below is the table of transformations, with thirteen identified in the present study. However, this list is not currently expected to be exhaustive.

Transformation	Meaning	Example
1	Denial with "no"	*-dar bola* *Ella le dio bola = Ella no le dio bola.*
2	Denial with "without"	*-hacer bolsa* *Ella hizo bolsa sus juguetes = Sin que haga bolsa tus juguetes.*
3	Passive voice	*-hacer bolsa* *María hizo mierda el auto = El auto fue hecho mierda por María.*
4	Replacement of C1 by Clitic	*-pegar la PLR* *María le pegó la PLR=* *María se LA pegó.*
5	Replacement of A1 with an accusatory clitic	*-hacer mierda* *Juan hizo mierda a María = Juan la hizo mierda*
6	Pluralization of the clitic	*-dejarle la cagada* *María le dejó =* *María les dejó la cagada.*
7	Interrogation	*-Dejar la escoba* *Juan dejó la escoba =* *¿Quién dejó la escoba?*
8	Conditional	*-Quedar pegado* *Juan quedó pegado =* *si Juan quedó pegado, María también.*
9	C1 with number inflection	*-Estar cagado* *Él está cagado =* *Ellos están cagados.*
10	C1 with gender inflection	*-hacer huevón* *Él lo hizo huevón=* *Él LA hizo huevona.*
11	C1 with diminutive	*-hacer tuto* *Ellos hicieron tuto =* *Ellos hicieron tutito.*
12	C with augmentatives	*-caer gordo* *Él le cayó gordo=* *Él le cayó gordísimo.*
13	Inclusion of a C1 modifier	*-Quedar la cagada* *Ayer quedó la cagada =* *Ayer quedó LA MANSA cagada.*

Table 7. Transformations of verbal idioms in Chilean Spanish.

At this point it is pertinent to clarify that transformations 11, 12 and 13 may correspond to the concept of de-automation (Zuluaga 2001; Mena Martínez 2003). This concept alludes to the phenomenon of deviation that a speaker makes from a phraseological unit of its canonical form, as exemplified by Mena Martínez (2003): *dar gato por euro* instead of *dar gato por liebre*. Generally, this is done for expressive, humorous reasons, etc., considered performative, which exceeds the objectives of this work. Nevertheless, the types of de-automation described in 11, 12 and 13 will be taken into account as transformations, since the semantic content of the VI is maintained.

Following the elaboration of the transformation glossary, we prepared the tables, with a fragment of the Class 1 transformation table presented below.

Verbo	C1	Ejemplo	Transformaciones											
			T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13
hacer	mierda.	Ellos hicieron mierda todo lo que provenía de sus enemigos	+	+	+	-	+	-	+	+	-	-	-	-	-
hacer	mierda.	La podadora hizo mierda todas las flores	+	+	+	-	+	-	+	+	-	-	-	-	-
hacer	mierda.	Ellos se harán mierda la ropa si siguen peleando	+	+	-	-	+	-	+	+	-	-	-	-	-
hacer	mierda.	Viene el aluvión y los carteles se harán mierda con la tormenta	+	+	-	-	-	-	+	+	-	-	-	-	-
hacer	mierda.	Pamela me hará mierda las uñas	+	+	-	-	+	-	+	+	-	-	-	-	-
hacer	mierda.	Las espinas me harán mierda la ropa	+	+	-	-	+	-	+	+	-	-	-	-	-

Table 8. Example of Class 1 idiom transformations.

In the previous table, the nuclear components of the idiom (V+C) are detailed first, followed by admitted transformations – which, as in the lexicon-grammar tables, is indicated with a + sign if the expression admits certain modification, or with a - if not. In the case of table 8, for example, the sentence *Ellos hicieron mierda todo lo que provenía de sus enemigos* takes:

Transformation	<OPT>
T1	Ellos **no** hicieron mierda todo lo que provenía de sus enemigos
T2	**Sin que** ellos hicieran mierda todo lo que provenía de sus enemigos.
T3	Todo lo que provenía de sus enemigos **fue hecho** mierda por ellos.
T5	Ellos **lo** hicieron mierda
T7	¿**Quiénes** hicieron mierda todo lo que provenía de sus enemigos?
T8	**Si** ellos hicieron mierda todo lo que provenía de sus enemigos…

Table 9. Example transformations.

As can be seen, the idiom described does not admit all transformations, so, for example, constructions such as (6) will be agrammatical.

(6) a. *Ellos los hicieron (T4)
b. *Ellos hicieron mierdas (T9)

Following these proposed exhaustive descriptions of VIs in Chilean Spanish, we proceed to computational modeling in NooJ.

3.2. Computation Work

Automatic detection of VIs in Chilean Spanish requires a formalization of the structures of each class. To this end, we used NooJ, a tool that has several utilities for the treatment of natural language and that can be organized from the hierarchy proposed by Chomsky and Schützenberger (1963). The program has the following resources:

- dictionaries: word lists with various types of linguistic information.
- morphological and derivative grammars: inflection and derivation models
- productive grammars: regular or graphic systems useful for the treatment of character chains with certain formal properties.
- syntactic grammars (.nog files): regular or graphic systems useful for the treatment of character strings formed by two or more lexical units, generally separated by blank spaces.

As mentioned, the word list in the dictionary includes information of various kinds, depending on the interests of the user (morphological, syntactic, lexical, semantic, etc.). In turn, lemmas can be associated with morphological models that allow automatic inflection. For example, a word such as *médico* is listed as follows:

médico,N+anim+hum+pm+FLX=NIÑO

with 'médico' as a noun ('N') that possesses the traits 'animate' ('+anim'), 'human' ('+hum') and 'medical professional' ('pm'), and has inflection ('FLX') corresponding to the NIÑO model, specified in the morphological grammar:

NIÑO = <E>/masc+sg | <B>s/masc+pl | <B>a/fem+sg | <B>as/fem+pl;

This procedure, which allows for automatic generation of grammars for *médico*, *médicos*, *médica*, *médicas*, is much more efficient and economical for inflection of numerous words (*médico*, *enfermero*, *farmacéutico*, etc.); in the case of Spanish verbs, which have an enormous wealth of inflections, the advantages of this type of grammar are indisputable (Bonino 2015).

At the same time, the program also has the possibility of elaborating grammars of various levels (depending on the context, finite states, etc.). For example, as can be seen in Silberztein (2016), a simple grammar for the English Noum Phrase (NP) can be schematized as follows:

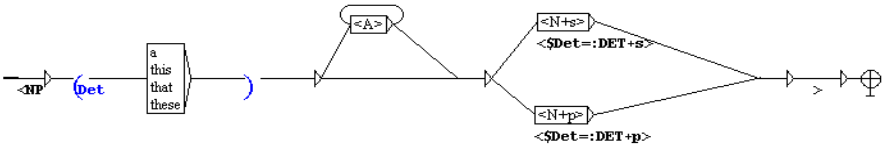


Figure 1. Example of a NooJ syntactic grammar for English (NP).

Here, the grammar begins with an initial state made up by a determinant - indicated in parentheses - which may include an adjective (<A>) and end in the nominal core (<N>). At the same time, the number matching rule is specified: if N has the singular trait ('+s'), then the number of the determinant will be singular; on the other hand, if N has the plural trait, the determinant will take that same trait. Thus, each expression containing this information will be labelled as NP (Noun Phrase) (<NP ... >). These grammars can operate both at the syntactic and morphological level and allow for automatic recognition and generation.

In sum, the procedures were: (i) elaboration of lexicon grammar; (ii) preparation of a dictionary that contemplates all the properties of each type of idiom.

3.2.1. Production of grammars and dictionaries

The grammar structures are exemplified in Table 10. Each elaborated grammar presents as basic cores: (i) the name of the idiom class; (ii) the unit V, which indicates the verbal nucleus of the expression; and (iii) the elements that make up the "fixed" part of the expression.

Example	V	PREPC	CONT	DETC	C
hacer mierda	Hacer	-	-	-	mierda
dejar la escoba	Dejar	-	-	la	escoba
atornillar al revés	Atornillar	-	al	-	revés
caer de cajón	Caer	de	-	-	cajón
agarrar para el chuleteo	Agarrar	para	-	el	chuleteo

Table 10. Example of composition of central elements of a Spanish verbal idiom.

The number following element C will change according to each class. For example, for class 1, it is called C1 and for class 2, C2 and so on. Grammars vary according to their constituent elements, with an example of class 1 shown in Figure 2.

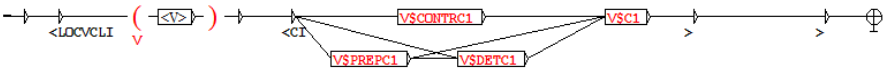


Figure 2. Lexical grammar for detecting class 1 phrase.

In this figure you can see the elements referred to above: at the beginning of the node, the name with which the expression 'LOCVLI' (Verbal Idiom Class 1); then, in brackets, variable V, with which the nuclear verb of the idiom will be detected; and, finally, in C1, the combinations of elements that can make up the fixed part of the idiom.

The grammar takes the information declared in the electronic dictionary, which contains the declared lemmas, as follows:

agarrar,V+FLX=AMAR+C1=papa+FXC
agachar,V+FLX=AMAR+DETC1=el+C1=moño+FXC

echar,V+FLX=AMAR+CONTRC1=al+C1=agua+FXC
caer,V+FLX=CAER+PREPC1=de+C1=cajón+FXC

The lemma begins with the verb of the idiom, followed by its inflection model (FLX=), and the invariable elements indicated by the +FXC command to flag it as a frozen expression.

With all this, it is possible to achieve detection even in those cases where the idiom has separate elements, as in *dejó Juan la mansa escoba*:

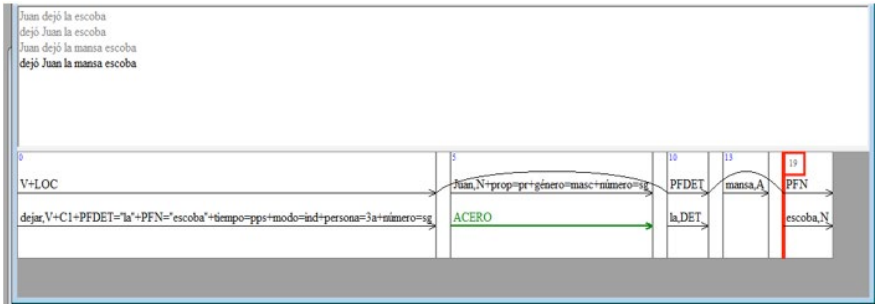


Figure 3. Example of class 1 phrase detection.

The detection "jumps" to unite the separate elements of the VI.

Next, Figure 4 shows the the syntactic grammar structure for class 2 VI.

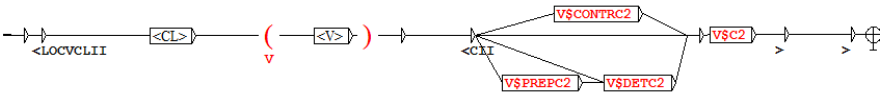


Figure 4. Lexical grammar for detecting class 2 idiom.

As can be seen, the only difference between this grammar and that of class 1 is the inclusion of the clitic pronoun before the verbal unit. This information must also be declared in the dictionary entries. Thus, at the time of detection, NooJ discriminates between elements that have a CL, as in this case, and those that do not, as in the previous case.

abrir,V+FLX=PARTIR+CL+C2=cancha+FXC
afirmar,V+FLX=AMAR+CL+DETC2=los+C2=pantalones+FXC
agarrar,V+FLX=AMAR+CL+CONTRC2=del+C2=moño+FXC
caer,V+FLX=CAER+CL+PREPC2=de+C2=maduro+FXC

An example of this type of detection is shown in Figure 5. Unlike the previous case, we have loaded class 1 and 2 dictionaries to demonstrate that this information allows NooJ to properly discriminate between one type of speech and the other.

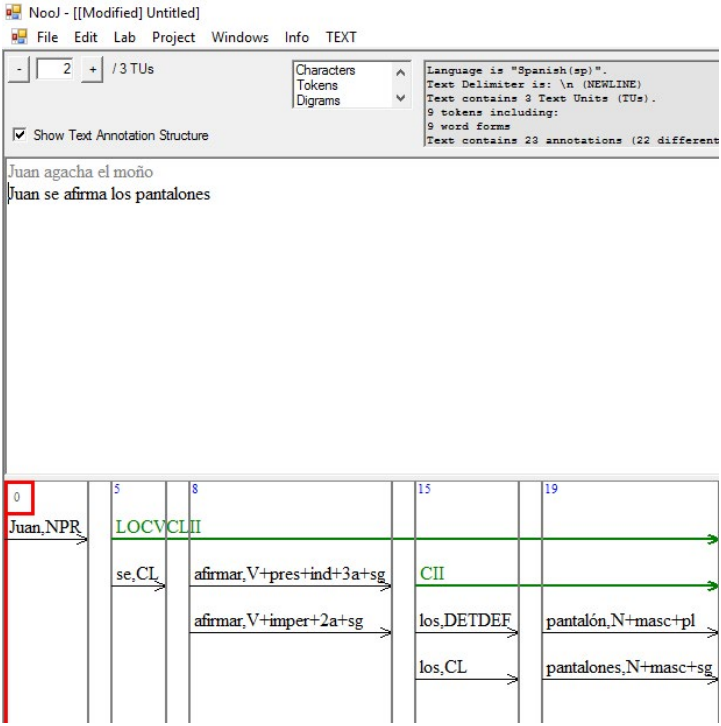


Figure 5. Example of class 2 idiom detection.

Even if a class 1 idiom supports a clitic pronoun in front of the verb, which would resemble class 2 grammar, NooJ performs the detection properly without mixing the information as provided by the grammars. As in the

previous example, Figure 6 shows the detection made using the dictionaries with the respective grammars of classes 1 and 2.

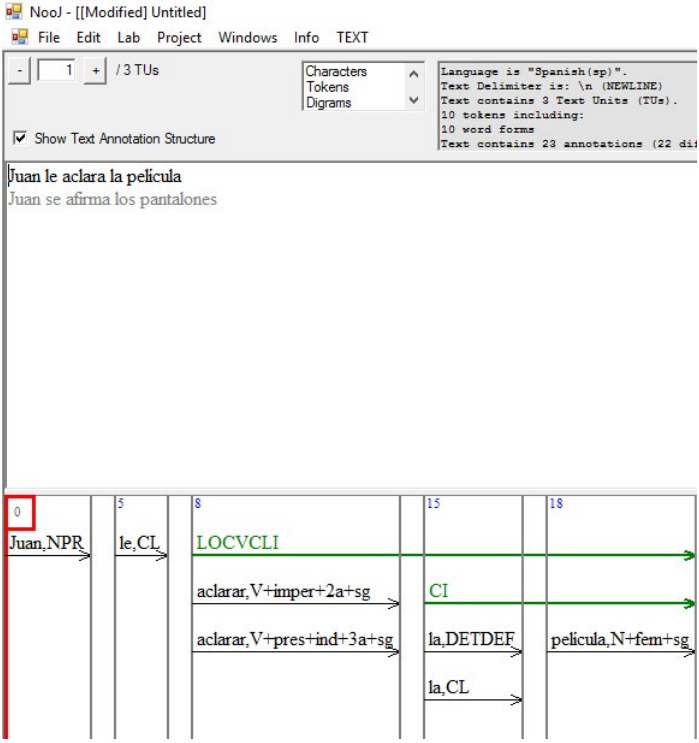
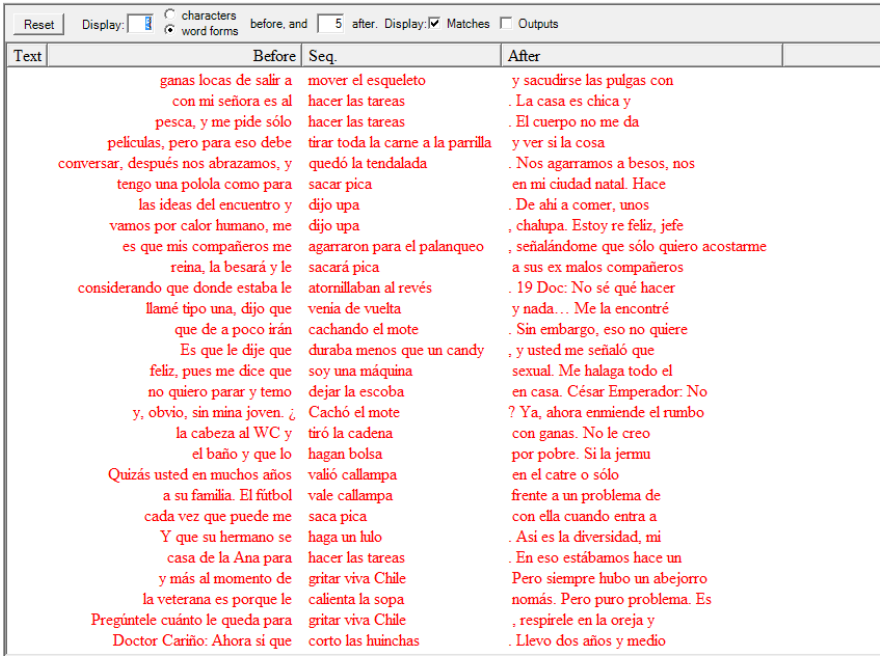


Figure 6. Example of detection for Class 1 idioms that support pronouns.

This methodology was tested in a corpus made of texts extracted from the Chilean press. In the following section, the results obtained are presented.

4. Results

To carry out the automatic detection, the dictionaries for each of the speech classes with their respective grammars were loaded into NooJ. Figure 7 shows an excerpt of the output made by the software.



Reset Display: 4 characters word forms before, and 5 after. Display: Matches Outputs

Text	Before	Seq.	After
	ganas locas de salir a	mover el esqueleto	y sacudirse las pulgas con
	con mi señora es al	hacer las tareas	. La casa es chica y
	pesca, y me pide sólo	hacer las tareas	. El cuerpo no me da
	películas, pero para eso debe	tirar toda la carne a la parrilla	y ver si la cosa
	conversar, después nos abrazamos, y	quedó la tendalada	. Nos agarramos a besos, nos
	tengo una polola como para	sacar pica	en mi ciudad natal. Hace
	las ideas del encuentro y	dijo upa	. De ahí a comer, unos
	vamos por calor humano, me	dijo upa	, chalupa. Estoy re feliz, jefe
	es que mis compañeros me	agarraron para el palanqueo	, señalándome que sólo quiero acostarme
	reina, la besará y le	sacará pica	a sus ex malos compañeros
	considerando que donde estaba le	atornillaban al revés	. 19 Doc: No sé qué hacer
	llamé tipo una, dijo que	venía de vuelta	y nada… Me la encontré
	que de a poco irán	cachando el mote	. Sin embargo, eso no quiere
	Es que le dije que	duraba menos que un candy	, y usted me señaló que
	feliz, pues me dice que	soy una máquina	sexual. Me halaga todo el
	no quiero parar y temo	dejar la escoba	en casa. César Emperador: No
	y, obvio, sin mina joven. ¿	Cachó el mote	? Ya, ahora enmiende el rumbo
	la cabeza al WC y	tiró la cadena	con ganas. No le creo
	el baño y que lo	hagan bolsa	por pobre. Si la jermu
	Quizás usted en muchos años	valió callampa	en el catre o sólo
	a su familia. El fútbol	vale callampa	frente a un problema de
	cada vez que puede me	saca pica	con ella cuando entra a
	Y que su hermano se	haga un lulo	. Así es la diversidad, mi
	casa de la Ana para	hacer las tareas	. En eso estábamos hace un
	y más al momento de	gritar viva Chile	Pero siempre hubo un abejorro
	la veterana es porque le	calienta la sopa	nomás. Pero puro problema. Es
	Pregúntele cuánto le queda para	gritar viva Chile	, respírele en la oreja y
	Doctor Cariño: Ahora sí que	corto las huinchas	. Llevo dos años y medio

Figure 7. Example of class 1 idiom detection.

The corpus used to test the automatic detection comes from national newspaper *La Cuarta*. Texts were extracted from "Ventanita sentimental", which publishes letters with various concerns sent to the newspaper, characterized by an informal register rich in verbal idioms. In total, the corpus contained 33,094 words, equivalent to more than one hundred letters and their respective answers.

The VIs included in the corpus were manually counted and compared with the NooJ output, which yielded the percentages of precision, coverage, and F measure. Idioms with higher enclitic verb structures were not taken into account (*sabiéndoselas por libro*), and will be addressed in future studies. The percentages achieved by the algorithm are listed in Table 11.

Results	Summary
Idioms by class	103
Detected well	95
Misdetected	3
Total detection	98
Coverage	92.23%
Accuracy	96.94%
Measure F	94.53%

Table 11. Total Detection Results.

The positive results allow us to verify that the formalization carried out via the methodological apparatus delivered by lexicon-grammar theory was adequate. However, due to the limited number of classes present in the corpus, it was not possible to validate the resources constructed for those absent. A future challenge necessary to study this type of expressions is to compile a corpus containing idioms of all classes.

The greatest number of idioms were concentrated in classes 1 and 2. This result is not unexpected, since over 50% of lemmas or entries in the DUECh belong to these classes; which indeed, reflects their diverse and common usage in Chilean Spanish. The least impressive results were for the coverage, which implies that future works will require grammars with more idiom properties.

Results	Class 1	Class 2	Class 4	Class 5
Idioms by class	61	47	**3**	**2**
Detected well	52	39	3	1
Misdetected	1	2	0	0
Total detection	53	41	3	1
Coverage	85.25%	82.98%	100%	50%
Accuracy	98.11%	95.12%	100%	100%
Measure F	91.23%	88.64%	100%	66.67%

Table 12. Detection results, by class.

Similarly, the qualities declared in the dictionaries should be reviewed in their inability to detect those idioms that were left out – 9, in the case of class 1; 8, in class 2; and 1, in class 5. Nevertheless, the results are satisfactory as a first approximation to the phenomenon of VIs in Chilean Spanish.

5. Conclusions

Our general objectives proposed were: (i) to describe the morphosyntactic properties and the argumentative structures projected by VIs in Chilean Spanish; and (ii) to detect Chilean Spanish VIs in natural language texts. These two main purposes guided the research work and, from the results shown above, were achieved. The methodological apparatus provided by lexicon-grammar theory was particularly effective for our descriptive work. The tables allowed for extraction of the qualities of the idioms and, at the same time, the formalization of their structures for subsequent computational modeling.

Interestingly, most idioms required an N0 + Human argument, and a large percentage of pronouns were dative (*Juan le da bola a María*). The "fixed" elements were shown mostly to determine the nominal or prepositional phrases. And, finally, the argument N1, whether introduced or not by prepositions, is not very frequent.

The descriptions were necessary for the formalization of these units into subsequent computational detection. Our model obtained positive detection results in all indicators, with precision, coverage, and F measure above 80%. The properties of the software facilitated this, with the option of a special command (+FXC) for the type of units to be detected.

Of the thirteen transformations contemplated in this descriptive study, three are recurrent across classes. Likewise, the high average idiom transformations admitted in each class shows their richness in syntactic-semantic and morphological properties. Despite being traditionally considered invariable, this vision must be reconsidered.

Although the objectives were generally achieved, some aspects remain outside the scope of this research and will have to be addressed in the future. As noted above, a project remaining is the elaboration of lexicon-grammar theory tables for group 2 idioms, which have the adverb "no"

in their dictionary entry; the elaboration of grammatical rules for detecting idioms with one or two enclitic pronouns in the verbal unit (hacérselo pebre); and the inclusion of more transformations in the table, and the formalization of these to enrich detection procedures. This may require a greater description of the constituent elements of the idioms, which translates into a more detailed description of the properties that these units possess.

Future descriptive work may be organized around the following pillars: (i) deepening linguistic descriptions in lexicon-grammar tables to improve automatic detection; (ii) extending the transformational possibilities of VIs in Chilean Spanish; and (iii) elaborating idiom tables for DUECh lemmas containing the negation adverb "no".

Future computational work may seek to: (a) expand the electronic VI dictionary; (b) enrich computer grammars; and (c) elaborate grammars for automatic generation of sentences containing predicate VIs. Likewise, to test these resources, a larger corpus must be elaborated.

References

ACADEMIA CHILENA DE LA LENGUA. (2010) *Diccionario de uso del español de Chile*. Santiago: MN Editorial Ltda.

BONINO, Rodolfo. (2015) "Una propuesta para el tratamiento de los enclíticos en NooJ." *Infosur Revista* 7, pp. 31-40.

CHOMSKY, Noam. (1957) *Syntactic structures*. Berlin: Mouton & Co.

CORPAS PASTOR, Gloria. (1996) *Manual de fraseología española*. Madrid: Gredos.

GROSS, Maurice. (1975) *Méthodes en syntaxe*. Paris: Hermann.

GROSS, Maurice. (1984) "Lexicon-grammar and the syntactic analysis of French." In: *Proceedings of the 10th International Conference on Computational Linguistics and 22nd anual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, pp. 275-282.

HARRIS, Zelig. (1970) *Papers in structural and transformational linguistics*. Dordrecht: Springer.

MENA MARTÍNEZ, Florentina. (2003) "En torno al concepto de desautomatización fraseológica: aspectos básicos." *Tonos Revista electrónica de estudios*

filológicos 5. Electronic version: <https://www.um.es/tonosdigital/znum5/estudios/H-Edesautomatizacion.htm>

ELIA, Annibale; Mario Monteleone & Federica Marano. (2011) "From the concept of transformation in Harris and Chomsky to the Lexique-grammaire of Maurice Gross." In: Kasevich, Vadim; Yuri Kleiner & Patrick Sériot (eds.) (2011). *History of Linguistics*. Amsterdam: John Benjamins, pp- 76-82.

PENADÉS, Inmaculada. (2012) *Gramática y semántica de las locuciones*. Alcalá de Henares: Servicio de Publicaciones de la Universidad de Alcalá.

PIVAUT, Laurent. (1989) *Verbes supports et vocabulaire technique: sport, musique et activités intelectuells*. Paris: University of Paris. Unpublished PhD dissertation.

SÁEZ, Leopoldo. (2012) "El léxico del dialecto chileno: *Diccionario de uso del español de Chile DUECh*." *Estudios filológicos* 49, pp. 137-155.

SILBERZTEIN, Max. (2003) *NooJ Manual*. Electronic version: <http://www.NooJ4nlp.net/>

SILBERZTEIN, Max. (2016) *Formalizing Natural Languages: The NooJ Approach*. London: Wiley Eds.

SILVAGNI, Federico. (2017) *Entre estados y eventos. Un estudio del aspecto interno del español*. Barcelona: Universidad Autónoma de Barcelona. Unpublished PhD dissertation.

TOLONE, Elsa. (2012) *Conversión de las tablas del Léxico-Gramática del francés en el léxico LGLex*. Ponencia presentada en el 2nd Argentinian Workshop on Natural Language Processing (WNLP'11), Córdoba, Argentina.

VIETRI, Simonetta. (2014) *Idiomatic Constructions in Italian. A Lexicon-Grammar Approach*. Amsterdam: John Benjamins BV.

ZULUAGA, Alberto. (1975). "La fijación fraseológica." *Thesaurus* 30, pp. 235-288.

ZULUAGA, Alberto. (2001) "Análisis y traducción de unidades fraseológicas desautomatizadas." *PhinN* 16, pp. 67-83.

BIONOTES

WALTER KOZA is PhD in Humanities and Arts with mention in Linguistics graduated at the Universidad Nacional de Rosario (Argentina). He is currently Assistan Professor at Pontificia Universidad Católica de Valparaíso

(Chile). His research areas are computational linguistics, lexicology and Spanish grammar.

NATALIA RIVAS is Teacher of Spanish and Communication and Bachelor in Language and Hispanic Literature, graduated at the Pontificia Universidad Católica de Valparaíso (Chile).