IMPLEMENTING EUROPEAN PORTUGUESE VERBAL IDIOMS IN A NATURAL LANGUAGE PROCESSING SYSTEM

Jorge Baptista
Univ. Algarve
L2F/INESC-ID
Lisboa
jbaptis@ualg.pt

Graça Fernandes
Rui Talhadas
Univ. Algarve
{gracitafernandes,
rtalhadas}@gmail.com

Francisco Dias
IST/Univ. Lisbon
L2F/INESC-ID
Lisboa
francisco.m.c.dias
@tecnico.ulisboa.pt

Nuno Mamede
IST/ Univ.
Lisbon
L2F/INESC-ID
Lisboa
Nuno.Mamede
@inesc-id.pt

Keywords: European Portuguese, verbal idioms, lexicon-grammar, rule-based parsing, natural language processing

Abstract

This paper is based on an extant *lexicon-grammar* of European Portuguese verbal idioms (*e.g.*, *deitar mãos à obra*, literally, 'to throw hands to the work', 'to start working'.). This a database containing about 2,400 expressions, along with all relevant information on the sentence structure, distributional constraints and transformational properties of these frozen sentences. In this paper, we present a solution to the integration of verbal idioms in a fully-fledged natural language processing system, and a preliminary evaluation using a small, manually annotated corpus.

1. INTRODUCTION

Verbal idioms (e.g. smb. *kill two birds with one stone*) can be defined as frozen sentences where the verb and at least one of its arguments are frozen together, and their overall meaning cannot be derived from the mere composition of the meanings of their individual elements, when used separately (Gross 1982, 1996; Cowie 1998). They constitute a large set of the lexicon and grammar of many languages, in the order of several thousands, though their frequency in texts is often very low. In fact, their occurrence it is highly dependent on text genre/type, being more common in oral than in written texts. The integration of verbal idioms in natural language processing (NLP) systems is relevant for an accurate semantic parsing. However, this integration is a challenge to NLP systems (Sag *et al.* 2002), as these idioms cannot be dealt with like other idioms, as frozen strings of words. In spite of their being semantically non-compositional, they do have syntactic structure, allowing inflection, insertions, several transformations and creative pragmatic reuse.

In this paper, we present a solution to the integration of verbal idioms in a fully-fledged natural language processing system, STRING (Mamede *et al.* 2012)[27]. This work is based on an

[27] string.l2f.inesc-id.pt

extant lexicon-grammar of European Portuguese verbal idioms (Baptista *et al.* 2004, 2005), containing about 2,400 expressions, e.g. *deitar mãos à obra* (lit: throw hand to work) 'start working'. Conceived in a tabular format, and organized in 10 main classes, according to the formal structure of the idioms, these tables present the verb and the frozen arguments of each idiom, along with the encoding of distributional constraints on free arguments and the structural changes (or transformations) the sentence can undergo (passive, pronominalization, etc.). Each idiom is also illustrated by an example.

In order to integrate the lexicon-grammar of verbal idioms in rule-based parsing module of the NLP system the following strategy was adopted: firstly, the general parsing rules are applied, so the frozen sentence is given a structure as any ordinary sentence; then, another set of rules extracts a (semantic) dependency (FIXED), based on the previous parse, and groups together the frozen elements of the idiom, while keeping intact the syntactic structure of the sentence; finally, FIXED dependency is then used to further calculate the sentence's semantics: for example, the semantic roles of the verb's free argument are to be extracted not from the information attached to the simple verb but to those of the verbal idiom; the part-whole semantic relations extraction is blocked; the verb sense statistical disambiguation module is prevented from acting, and so on. A script automatically reads the tabular format and converts it into the syntax of the rule-based parser for the extraction of the FIXED dependency. To assess the conversion process, the set of rules was applied to the examples provided in the lexicon-grammar. A small percentage of errors were detected and some rules were manually adjusted. Most errors, though, were due to incorrect part-of-speech (POS) tagging.

To evaluate the system's new module, sentences including all the key-elements of each idiom were extracted. Then, a team of linguists manually annotated a representative sample, taken from a freely available corpus, in order to build a golden standard. Then the sentences were parsed and results were automatically compared to the golden standard. A detailed error analysis is also briefly presented.

This paper is structured as follows: Section 2 briefly presents the lexicon-grammar of European Portuguese verbal idioms. Section 3 described the construction of an annotated reference corpus of verbal idioms. Section 4 describes the implementation of the verbal idioms' identification module in the STRING system, while Section 5 presents and briefly discusses the evaluation of the module using the annotated corpus. Finally, section 6 concludes the paper and suggests future work.

2. A LEXICON-GRAMMAR OF EUROPEAN PORTUGUESE VERBAL IDIOMS

Frozen sentences or *verbal idioms* can be defined (Gross 1982, 1996) as sentences where the verb and at least one of its argument noun-phrases are frozen together, that is, they are distributionally constraint. In a free sentence, the meaning is determined from the individual meaning of the elements in the construction, but, the meaning of the frozen sentence is non-compositional, *i.e.* it cannot be directly calculated from the meaning that the component elements may present when used separately. For example, in *brincar com o fogo* (lit: "to play with fire"), neither the verb *brincar* 'to play' nor the noun *fogo* 'fire' can be substituted by any other word, unless a change is seen in the sentence's overall meaning: 'to do dangerous, risky things'.

The structure of these type of sentences is similar to that of free sentences, but syntactic and distributional constraints cannot be calculated or predicted from the formal properties that the elements of expression may have when constructed separately. For example:

(1) *O Pedro saiu* (*do armário* + °*da sala* + °*da loja*)

literally: 'Peter left/exited [from] the closet/room/shop'

'to assume one's (homo)sexuality'

(2) *A Maria fechou-se em* (*copas* + **paus* + **espadas*)

literally: 'Mary closed herself in hearts/clubs/spears'

'to be silent, not to disclose information'

In (1), the verb *sair* 'leave/exit' requires a locative-source complement, which the noun *armário* 'closet' can fill in, but in the frozen idiomatic interpretation only this noun can occur, otherwise the sentence's interpretation becomes literal (signaled by '°'). In (2), the noun *copas* 'hearts' cannot be replaced by any other deck of cards, and it does not corresponds to the ordinary distribution of the reflexive use of the verb *fechar* 'close'.

The European Portuguese verbal idioms were classified into 10 formal classes according to their structure and distributional constraints. The theoretical and methodological framework here adopted is the Lexicon-Grammar (M. Gross 1982, 1996), based on the Harrissian transformational operator-grammar (Z. S. Harris 1991). Table 1 shows the structure of each class studied in this work[28].

Class	Structure	Example	Count
C1	N_0 *V* C_1	*O Pedro bateu o pé*	
		to beat the foot 'to refuse to do smthg'	491
CDN	N_0 *V* (*C de N*)$_1$	*O Pedro aprendeu o bê-á-bá da gramática*	
		'to learn the basic concepts of smthg'	45
CAN	N_0 *V* (*C de N*)$_1$ = C_1 *a* N_2	*O Pedro cortou as asas* (*do João* = *ao João*)	
		to cut the wings to/of smbd	
		'to prevent smb from acting freely'	175
CNP2	N_0 *V N1 Prep*$_2$ C_2	*O Pedro não tirava a Ana da cabeça*	
		not to take smbd/smthg from the head	
		'think continuously on smb/smthg'	172
C1PN	N_0 *V* C_1 *Prep*$_2$ N_2	*A Rita cravou as unhas na fortuna do João*	
		to dig the nails into smthg	
		'to acquire/steal smthg'	233
C1P2	N_0 *V* C_1 *Prep* C_2	*O Pedro deu o dito pelo não dito*	
		to give the said for the non-said	
		'to change one's opinion, not to be true to one's word'	288
CPPN	N_0 *V* C_1 *Prep* C_2 *Prep* C_3	*O Pedro deitou fora o bebé com a água do banho*	
		to throw away the bay the bath water	26

[28] The code for each class is purely conventional; *N* and *C* stand for noun phrases; *N* is a free constituent and *C* is frozen noun phrase; N_0 is the subject, N_1 and N_2 the first and second complement; *V* is the verb and *Prep* a preposition. These codes and the defined classes are the same as the ones proposed initially by M. Gross (1982, 1996). The example (in *italics*) is followed by a literal translation and gloss.

		'to lose the important along with the non important'	
CPP	*N_0 V Prep C_1 Prep C_2*	*O Pedro deu com o nariz na porta*	
		to hit with the nose on the door	
		'to go somewhere in vain'	201
CP1	*N_0 V Prep C_1*	*O Pedro passou pelas brasas*	
		go through the embers	
		'take a nap'	703
CPN	*N_0 V Prep (C de N)$_1$*	*O Pedro não chega aos calcanhares do João*	
		not to reach the heels of smb	
		'not to be a match for smb'	95
		Total	2417

Table 1. Classification of Frozen Sentences of the European Portuguese.

The relevant linguistic information has been encoded in binary matrices, where each line corresponds to a lexical entry, and the columns contain either the lexical elements of the expression or the signs '+' and '-' to encode the relevant linguistic properties it presents. Properties include the distributional constraints (human/non-human) of the free argument slots and certain transformations, such as *Passive* or obligatory complement permutation. For this paper, the most relevant properties are:

(a) *intrinsic reflexive constructions* (noted *V se*), where the verb presents a reflexive pronoun that can not be derived by pronominalizing a free noun phrase, *e.g.*

(3) *O Pedro pôs-se em bicos dos pés* (lit: 'Pedro put himself on the tip of the feet')

'to pretend exaggerated self-importance'

cp. **O Pedro pôs o João em bicos dos pés* (lit: 'Pedro put João on the tip of the feet');

(b) *obligatory negation constructions* (noted *NegObl*), where an expression can only be used in the negative:

(4) *O Pedro não dá para as encomendas* (lit: 'Pedro does not give to the requests')

'to be unable to meet the demands'

cp. **O Pedro dá para as encomendas.*

3. BUILDING AN ANNOTATED REFERENCE CORPUS OF VERBAL IDIOMS

A small corpus was prepared to evaluate the performance of the parser. This evaluation corpus was retrieved from the CETEMPúblico corpus (Rocha & Santos 2000), a publicly available corpus of journalistic text (dated from 1991 to 1998). The corpus is made of unrelated text extracts. For this paper only the first fragment of the corpus was used, featuring over 12 million words. For the evaluation corpus, the selection was made retrieving the extracts containing simultaneously the idioms' main verb and their frozen elements (prepositions and frozen NP head nouns) in the manner described below.

The UNITEX 3.0 linguistic development platform (Paumier 2003, 2014)[29] was used to retrieved these expressions from the CETEMPúblico fragment. UNITEX is based on finite-state automata (FSA) technology and it is able to intersect the linguistic data encoded in the lexicon-grammar matrices with finite-state transducers, which are then used to find patterns in texts and modify them, as well as to extract matching textual units (sentences) from larger texts. First, reference graphs are built, one for each class of idioms. Fig. 1 illustrates the reference graph for class **CP1**.

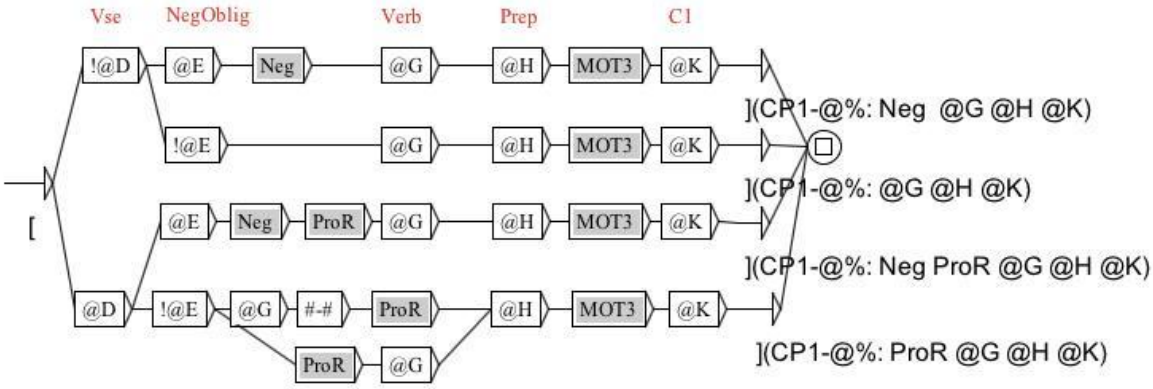

Figure 1. Reference graph for class CP1 (*e.g.* O Pedro passou pelas brasas, lit: Pedro passed through/over the burning coals 'Pedro took a nap')

These graphs refer to the data in the matrices by way of a set of variables @X, where 'X' stands for the corresponding column in the matrix. The variables can also be used as switches, allowing a path to be followed if the given cell contains a plus '+' sign or collapsing the path at that point if the cell contains a minus '-' sign. Switches can also be denied '!@X', which as the opposite effect. The graph reads as follows (symbols in red are just comments, to help reading the graph, and are not taken into account): variables @D and @E refer, respectively, to columns D and E of the lexicon-grammar matrix of idioms' class **CP1**, where the intrinsic reflexive and the obligatory negation properties, respectively, have been encoded. They function as switches. There can be four combinations of these property-value pairs, each with the corresponding path in the graph. Sub-graphs are represented by grey nodes: Neg includes the most common negation adverbs, ProR lists the reflexive pronouns and MOT3 is an insertion window from 0 to 3 words. Variables @G, @H and @K stand for the verb, the preposition and the head noun of the frozen complement. In the output, the matched is delimited by square brackets, while a simple tag is added after it, consisting of the class code, the line number in the matrix (variable @%) and the key elements of the idiom. An equivalent graph is hand-built for each class. The reference graph is then intersected with its respective matrix, reading it line-by-line and building an FST for each idiom (see Fig. 2). A general graph is then produced containing all the FSTs for that class.

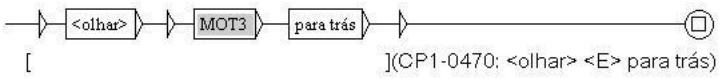

Figure 2. Finite-state transducer automatically built from the CP1 lexicon-grammar matrix for the idiom *olhar para trás* 'to look back (=into the past)' .

[29] www-igm.univ-mlv.fr/~unitex

The resulting transducer can thus be used to retrieve, delimit and tag the matching sequences from the corpus. Table 2 shows the breakdown of the raw 562 matching sequences per class and their distribution. There were 270 different expressions receiving different tags and their frequency ranged from a single instance (or *hapaxes*; 159 unique instances) to 42 (*v.g.* CAN-0045:<*chamar*> <DET> *atenção*); the difference between the number of matches and counts per class is due to the fact that some idioms could correspond to more than one lexical entry:

Class	**C1**	**CDN**	**CAN**	**CNP2**	**C1PN**	**C1P2**	**CPPN**	**CPP**	**CP1**	**CPN**	**Total**			
Matches	170	170	84	24	42	12	0	9	98	20	629			
Bin	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**12**	**25**	**42**	**Total**
Count	159	48	28	7	8	4	1	5	2	3	3	1	1	562

Table 2. Breakdown of the matching sequences per class and per frequency bins

The text units (sentences) matched by UNITEX were then manually annotated by 3 linguists, all well experienced in corpus annotation tasks and very familiar with the concepts involved in the study of idioms. A set of guidelines with examples was also provided and the annotation process was carried out independently. Each sentence could be tagged as follows:

tag	**description**
fixed :	the matched string corresponds to the targeted idiom;
fixed-different :	the matched string corresponds to an idiom but not to the targeted idiom (lexical elements involved can be slightly different, and/or the class is different); this happens because some idioms share the same lexical items, and usually the system chooses the longest match;
literal :	the matched string corresponds to the targeted idiom (eventually, it only partially corresponds), but this sequence of of words is being used in a literal, non-idiomatic way (e.g. *O Pedro deu um berro* 'Pedro gave a scream (='yell') vs. *O portátil deu o berro* 'The laptop broke');
false-positive :	the matched string contains elements of an idiom, but in that context it has nothing to do with the target idiom;
PoS-error :	the matched string includes an incorrectly PoS-tagged item;
other :	other problems not mentioned above.

Table 3 shows the distribution of the tags by the 3 annotators. The inter-annotator agreement was measured using ReCal 0.1 Alpha for 3+ Coders[30]. Table 4 shows the percentage of agreement between annotators:

tag	**Annotator 1**	**Annotator 2**	**Annotator 3**
other	0	15	33
fixed	399	373	398
fixed-different	37	41	55
false-positive	103	4	32
literal	63	158	86
POS-error	27	37	25

Table 3. Distribution of the tags by the annotators.

[30] dfreelon.org/recal/recal3.php

Average Pairwise Percent Agreement			
average	**annotators 1 & 3**	**annotators 1 & 2**	**annotators 2 & 3**
0.67	0.70	0.66	0.65
Average Pairwise Cohen's Kappa			
average	**annotators 1 & 3**	**annotators 1 & 2**	**annotators 2 & 3**
0.435	0.471	0.432	0.403

Table 4. Inter-annotator agreement.

The inter-annotator agreement can be considered as only moderate and it is similar pairwise; the Fleiss Kappa is 0.431, for an observed agreement of 0.671 and an expected agreement of 0.432, which is usually interpreted as 'fair'; the average pairwise Cohen Kappa is also deemed as 'fair'. This level of agreement may due to the difficulty of the task, particularly with the distinction between a literal and a figurative use, and the tag 'false-positive', which are very differently distributed among annotators, perhaps in a significant way. There were also an important number of cases where the annotators were unsure of the tag (and selected 'other').

A golden standard was established based on the consensual or most voted tag. In the end, 'PoS-error' and 'other' sentences were removed. The resulting corpus contains 602 sentences, 432 positive instances (noted 'fixed' and 'fixed-different') and 170 negative instances (false-positives and literal). The positive instances contain the target expression delimited and annotated for its class, ID number, key lexical elements (verb and frozen prepositions and head nouns). The negative instances contain just the tag 'non-fixed'.

4. IMPLEMENTING VERBAL IDIOMS IN THE XIP FORMALISM

Dealing with idioms in natural language processing systems is difficult, among other reasons, because their architecture must be conceived in such a way that it should not preclude the processing of both free word combinations and these, more constraint, expressions. On the other hand, many idioms do have syntactic structure, and can undergo several types of formal variation, thus making them hard to identify in a strictly string pattern-matching approach. Furthermore, many of these expressions are ambiguous between a literal (non-idiomatic) and figurative, non-compositional (idiomatic) use, depending of many linguistic and extra-linguistic factors.

In this section, we present the way European Portuguese verbal idioms have been integrated in STRING (Mamede *et al.* 2012), a hybrid, statistical and rule-based, fully fledged natural language processing system. STRING has a modular structure, shown in Fig. 3, that performs all the basic NLP tasks in four main steps: (1) preprocessing (tokenization, sentence splitting and lexical analysis; (2) rule-based and context-depending PoS disambiguation, MWE detection and context-depending contraction splitting; statistical PoS disambiguation; and (4) parsing, using the rule-based XIP parser (Xerox Incremental Parser, Ait-Moktar *et al.* 2002). Additional external modules operate on the output of XIP to perform other NLP-specific tasks, such as anaphora resolution, time normalization or slot filling.

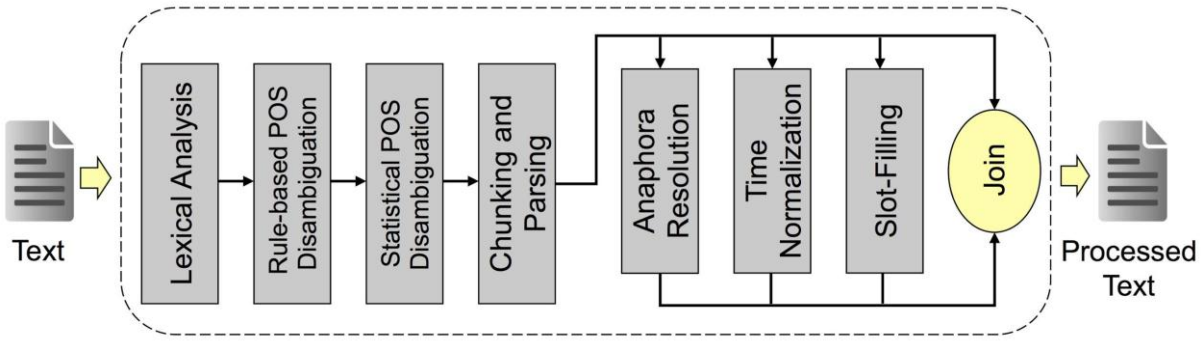

Figure 3. STRING architecture (Mamede *et al.* 2012)

The rule-based Portuguese grammar used by XIP was developed by the L2F (INESC-ID Lisboa) team in collaboration with Xerox and, besides adding to the lexicon some syntactic-semantic information required for the parsing stage, its processing consists of two main steps: (1) *chunking* (or *shallow parsing*), that is, the delimitation of elementary constituents (or chunks), such as NP, PP, etc.; and (2) *deep parsing*, that is, the extraction of syntactic dependencies between the chunks heads e.g. SUBJect, MODifier, CDIR (direct complement), *etc.*

Next, the general strategy for processing verbal idioms in STRING is laid out in order to better understand the automatic conversion process of the idioms' lexicon grammar matrices into the formalism XIP dependency rules.

Verbal idioms are identified in STRING by means of a dependency FIXED linking the key elements of the structure (the main verb, the prepositions and frozen head nouns). Below is an example of a sentence, the corresponding chunking and the (relevant) dependencies the parser extracted, including the FIXED dependency that identifies the idiom *medir as palavras*, literally 'to measure one's words', 'to be prudent when speaking':

```
SUBJ_PRE(mediu,Pedro)
CDIR_POST(mediu,palavras)
FIXED(mediu,palavras)
0>TOP{NP{O Pedro} VF{mediu} NP{as palavras} .}
```

This FIXED dependency is extracted based on the linguistic information already available in the system and the syntactic structure already calculated at the time the idioms identification module is applied. In this example, the verb and the direct object have been correctly PoS-tagged, the corresponding verb and noun phrase chunks properly identified, and a direct complement (CDIR) dependency between the verb and the head of the noun phrase has already been extracted by the time the idioms identification module comes into play. A dependency rule in this module then identifies the idiom by extracting the FIXED dependency:

```
if (  VDOMAIN(#?,#2[lemma:medir]) &
         CDIR[post](#2,#3[surface:palavras]) )
     FIXED(#2,#3)
```

This rule, which has been simplified here for clarity purposes, reads as follows: first a set of conditions (if) is stated, involving several variables; variables are identified by *#n* ; the first condition captures a verbal chain formed by a string of auxiliary verbs and a main verb, whose lemma is *medir* 'measure'; the next condition verifies if there is a CDIR dependency between the main verb and the noun *palavras*; finally, the dependency is extracted between the verb and the adverb.

This rule formalism requires that the information encoded in the lexicon-grammar matrices be converted into the XIP syntax. Writing rules for XIP grammar is a hard and time-consuming task. Filling a table is a much simpler task to a human than writing rules on XIP syntax. A faster approach for rule implementation could be to write an intermediary representation of the patterns. Then, this intermediary representation could be automatically converted into XIP rules. In order to represent these patterns, a table with the expected elements of the syntactic dependency links for each frozen expression was built. In this table each column represents a given element of a chunk of the syntactic dependencies structure. Any number of chunks can be concatenated to the table. An element of a chunk can be represented in such table in five different ways, which can be combined, as presented in Table 5:

Identification of the element	**Example**	**Description**
Surface	*andar*	Accepts only a word with *andar* 'walk' as surface
Lemma	<andar>	Accepts any inflection of lemma *andar* 'walk'
PoS tag	<DET>	Accepts any determinant (DET is the POS tag for determinant)
Dependency relation	MOD:<andar>	Word with lemma *and*ar 'walk' must be the chunk head which is a modifier (identified by the prefix MOD)
Grammatical features	<DET:ms>	Accepts any determinant with traces masculine and singular

Table 5. Intermediate representation of the information present in the lexicon-grammar.

The presence of a given element can be also negated using the identifier <E>. By default all modifiers are linked to the previous NP or PP (as in the case of *de N* 'of N' so-called determinative complements). That, however, is not always true due to the well-known PP-attachment ambiguity problem. To avoid PP-attachment ambiguity, a special flag (AttachV) can be placed to link a modifier directly to the verb of the sentence. Table 6 shows one entry and some of the columns of the lexicon-grammar for class CP1 and the idiom *ir desta para melhor* (lit: to go from this one to a better one, 'to die'):

SubjHum	Verb	Prep1	Det1	Mod1	Prep2	Det2	Mod2	AttachV2
+	<ir>	de	<E>	esta	para	<E>	melhor	+

Table 6. An idiom entry in the lexicon-grammar.

The resulting FIXED dependency is slightly different from the previous example:

```
FIXED_NORMALIZED(morrer,foi,esta,melhor)
```

In this case, the feature _NORMALIZED is added to the dependency, and a conventional classifying word, *morrer* 'to die' is added to its argument set. This information is encoded in the lexicon-grammar for certain types of semantic predicates that are relevant for the extraction of relations having to do with the major events in peoples biography, such birth and death, marriage or divorce, family ties, etc.

Any number of modifiers can be added in any order to the table. This allows any of the 10 classes of Portuguese frozen expression to be represented in the same table, by filling different elements for each class.

The system was first tested on the illustrative examples provided with the lexicon-grammar entries. These are artificial examples, where the idiom is expressed in the barest of forms, usually in the past or present tense, without any modifiers or adjuncts. Next, we describe the results of this preliminary test and the solutions adopted for the problems found. Table 7 presents the number of sentences used for testing and the errors found for each class of fixed expressions.

Class	Entries	Errors	% Error
C1	491	10	0.02
CDN	45	4	0.11
CAN	173	7	0.05
CNP2	172	22	0.13
C1PN	233	17	0.08
C1P2	288	26	0.09
CPPN	26	2	0.12
CPP	201	16	0.08
CP1	703	53	0.08
CPN	95	6	0.07
Total	2,417	163	0.07

Table 7. Error rate in the parsing of the lexicon-grammar's examples.

Overall, the percentage of correct cases is 93%, which constitutes a relatively high accuracy. It should be noted that the idioms identification module of the XIP parser Portuguese grammar is executed at a very late stage of parsing, thus it is hindered by all the errors that have accumulated in the pipeline up until then.

In the following, we investigate the main causes for the cases where the parser was unable to detect and extract the FIXED dependency. A preliminary classification of the error types was used to guide the assessment of this phase. Errors can be divided into 4 different types: (a) incorrect part-of-speech (PoS) tagging; (b) incorrect dependency extracted; (c) incorrect chunking; and (d) lexical gaps. Table 8 presents the breakdown of the types of errors, which are then discussed below.

Type of Error	Count
incorrect POS	62
incorrect dependency	47
incorrect chunking	41
lexical gaps	12
Total	162

Table 8. Distribution of the error types.

(a) incorrect PoS tagging

The STRING system automatically assigns a PoS tag to each token, using both a rule-based and a statistical PoS tagger. In average, the tagger achieves a state-of-the-art precision of 98%.

Naturally, like any other part of a text, frozen sentences can feature PoS-tagging errors. For example, in (1):

(1) *O João meteu baixa* (lit: 'João put a sick-leave') 'João called off sick'

the word *baixa* 'sick-leave' should have been classified as a noun, but it was tagged as an adjective, *baixa* 'short'. After this, the chunking module of the parser builds an AP chunk (adjectival phrase) instead of a NP, thus precluding the direct complement dependency (CDIR) from being extracted. As the frozen sentence dependency rule is built upon the extraction of the CDIR, the system fails to extract the FIXED dependency. This is the most frequently occurring error in the examples here tested. One of the possible solutions for this is to improve the rule-based PoS disambiguation, using contextual rules whenever the word combination is PoS-unambiguous.

(b) incorrect dependency extracted

The parsing performed by the STRING extract syntactic dependencies between the constituents of the sentence, using the syntactic and semantic features that have been added to the word in the lexicons. Since, at this time, word-sense disambiguation is only performed by the system regarding verbs (Travanca 2013, Suíssas 2014), certain ambiguous words can lead to an error when extracting dependencies. Sentence (2) is a good example of this:

(2) *O João tirou partido da notícia* 'João took advantage from that news'

Since the PoS-tagging and the chunking are correct, the relation CDIR should have been extracted between the verb *tirar* 'take' and the noun *partido* 'advantage'. However, *partido* is an ambiguous noun: besides this use as an abstract-uncountable noun, it also has the feature *group-of-things*, for it can be a human collective (a political party, for example). This feature triggers a QUANTD dependency to be extracted instead. This would correspond to an interpretation where *partido* would function as a quantifying determiner, like group of things, *e.g.*, as in the phrase *um partido de pessoas de esquerda* 'a party of left-wing people'. Therefore, *notícia* becomes instead the direct complement of *tirar*, which is incorrect. In order to capture this expression the noun *partido* should have been word-sense disambiguated first, removing its feature *group-of-things*.

(c) incorrect chunking tree

For each sentence being analyzed, STRING presents the resulting chunking tree. Errors may occur in this previous parsing task, which can lead to failure in recognizing the idiom. Consider sentence (3):

(3) *O João lava daí as suas mãos* (lit: João washes from there his hands)

'to wash one's hands (like Pilatus)'

In the lexicon-grammar, the idiom constituents are presented in their canonical order, which would be **O João lava as suas mãos daí*. However, in this idiom, the direct complement and the prepositional phrase are obligatorily reversed. This leads to the following chunking:

```
0>TOP{NP{O João} VF{lava} PP{de aí NP{as suas mãos}} .}
```

where a single PP is chunked incorrectly, integrating the adverb *aí* 'there' and the NP *as suas mãos* 'his hands'. Since the identification of the idiom depends on the previous correct identification of the two constituents, the CDIR {*as suas mãos*} and the MODifier PP {*de aí*}, the system fails to capture it. Chunking rules are a key element in the parsing process and are not to

be changed lightly, so in this cases, a manual rule has to be written, based on the idiom's (incorrect) chunking.

(d) lexical gaps

In spite of its large lexical coverage, STRING lexicons, especially multiword expressions, may still have some *lacunae*. The compound preposition *por debaixo* de in (4) was an example of such missing items:

(4) *O Pedro passou o dinheiro por debaixo do pano*.

(lit.: 'Pedro passed the money under the rug') 'Pedro made bribe'

In this case, it suffices to add the missing item to the lexicon, and the idiom can be properly identified. Other such cases involved compound nouns like *quadratura do círculo* 'circle's quadrature', (*discutir*) *o sexo dos anjos* '(to discuss) the angels' sex'.

During error analysis, some errors in the lexicon-grammar were also spotted (and corrected). For example, the codes for possessive determiners, as in *gastar a minha saliva* (lit: 'spend my salive') 'talking idly', were missing and had to be introduced in the appropriate cells, in order to capture such expressions. In other cases, a more general intervention in the grammar was required. For example, in the idiom *não ligar nenhuma a N* 'not minding nothing-fem.sg. to smthg' the indefinite *nenhuma* 'none' is obligatorily in the feminine-singular form and the verb prepositional (free) complement is introduced by *a* 'to'. However, a partitive quantifying determiner dependency was being extracted between *nenhuma* 'none' and the head noun of the PP. This should only happen in cases like *nenhum dos livros* 'none of the book', where gender agreement is required between the indefinite and the head of the PP and the preposition must always be *de* 'of'. In this case, the rule for extracting partitive quantifying determiner was made more precise.

5. EVALUATION AND DISCUSSION

The sample of 602 sentences that constitute our evaluation corpus were then processed by STRING and compared against the golden standard. Results are presented in Table 9:

Run	TP	FP	TN	FN	Precision	Recall	F-measure
1st	81	8	165	348	0.91	0.19	0.31
2nd	174	44	131	253	0.79	0.41	0.54

Table 9. Evaluation results.

The initial recall (1st run) was very low, though precision was high. We have undertaken the error analysis to understand the reasons for these suboptimal results. The main reason seems to be the fact that rules require a human subject to be explicit in the sentence. However, subject drop is a frequent phenomenon in Portuguese, so this condition should not be part of the rules (at least not at a first step). The same happens with several free complements (both NP and PP). After reformulating the script that generates the rules, when the system ignores the distributional constraints (2nd run) recall, though low improved to the double, while precision dropped 0.12.

Another reason for low results is the variation of the preposition *a/para* 'to', very common in the corpus, but missing altogether in the lexicon-grammar. The later preposition is also more common in Brazilian Portuguese. The annotators noticed a large number of sentences from the Brazilian Portuguese variant in the evaluation corpus, and, though there are significant differences between the idioms of each variant (Baptista 2008), this did not seem to have a significant impact in the results.

6. CONCLUSION AND FUTURE WORK

This paper presented the complex issues involved in the integration of a large-sized lexicon-grammar of European Portuguese verbal idioms into a natural language processing system, STRING, evaluating the resulting identification module of the system's parser on a manually annotated corpus. While precision was high, recall is pretty low: these results were hindered by the too restrictive rules, imposing the verification of distributional constraints on subject and complement, when these elements (especially the subject) can often be omitted in Portuguese. This is the single most important task to be completed in the near future. The evaluation has also shown that in the pipeline structure of the system, as errors tend to accumulate, some idioms are difficult to capture due to errors in previous processing stages, especially in PoS disambiguation and chunking. Some errors were also due to the incomplete word-sense disambiguation.

The paper showed that it is possible to identify idioms in texts, with a rule-based approach, based on lexical information and on the syntactic structure, as this has been previously calculated by the natural language processing chain, using the grammar of the general language. Thus, the system uses (and maintains) the syntactic structure of idioms (only a few cases present divergent syntactic constraints and require manual encoding).

In this way, the linguistic description is also kept apart from the processing issues. On one hand, this will benefit the continuing process of collecting the rich inventory of verbal idioms in the language. On the other hand, it will improve the semantic processing of Portuguese texts, as these meaning units are now better identified, affecting such disparate tasks as word-sense disambiguation, coreference resolution and semantic role labeling.

Acknowledgments

Research for this paper was partially funded by Portuguese national funds through Fundação para a Ciência e a Tecnologia, under project PEst-OE/EEI/LA0021/2014.

References

AIT-MOKHTAR, S; CHANOD, J.; ROUX, C. 2001. Robustness beyond shallowness: incremental dependency parsing. *Natural Language Engineering* 8 – 2/3: pp. 121-144.

BAPTISTA, J. 2004. Compositional vs. Frozen Sequences. Laporte, Eric; Ting Au-Chen, (eds). Proceedings of the Lexicon-Grammar Workshop. Beijing 14-18 de Outubro de 2004. *Journal of Applied Linguistics*, Special Issue on Lexicon-Grammar. Papers presented at the Lexicon-Grammar Workshop, pp. 81-93 (Chinese version).

BAPTISTA, Jorge; MAMEDE, Nuno; MARKOV, Ilia. 2014. Integrating verbal idioms into an NLP system. In: Jorge Baptista, Nuno Mamede, Sara Candeias, Ivandré Paraboni, Thiago Pardo, Maria das Graças Volpe Nunes, (Eds.). *Computational Processing of the Portuguese Language.* 11th International Conference

PROPOR'2014, São Carlos – SP, Brazil, October 8-10, 2014. Proceedings. pp. 251-256. *Lecture Notes in Computer Science / Lecture Notes in Artificial Intelligence* 8775: Berlin: Springer.

BAPTISTA, Jorge; CORREIA, Anabela; FERNANDES, Graça, 2004. Frozen Sentences of Portuguese: Formal Descriptions for NLP. *Workshop on Multiword Expressions: Integrating Processing, International Conference of the European Chapter of the Association for Computational Linguistics*, Barcelona (Spain), July 26, 2004. – ACL: Barcelona, pp. 72-79.

BAPTISTA, Jorge; CORREIA, Anabela; FERNANDES, Graça, 2005. Léxico Gramática das Frases Fixas do Portugués Europeo. *Cadernos de Fraseoloxía Galega* 7, Santiago de Compostela, Xunta de Galicia/Centro Ramón Piñero para a Investigación en Humanidades, pp. 41-53.

COWIE A., 1998. *Phraseology. Theory, analysis, and applications*. Oxford: Oxford University Press.

GROSS, M. 1982. Une classification des phrase "figées" du français. *Revue Québécoise de Linguistique* 11-2: pp. 151-185.

GROSS, M. 1996. Lexicon-Grammar. *Concise Encyclopedia of Syntactic Theories*. Cambridge. Pergamon. pp. 244-258.

HARRIS, Zellig S. 1991. *A Theory of Language and Information: A Mathematical Approach*. Oxford: Clarendon Press.

MAMEDE, Nuno; BAPTISTA, Jorge; DINIZ, Cláudio; CABARRÃO, Vera. 2012 STRING – An Hybrid Statistical and Rule-Based Natural Language Processing Chain for Portuguese. in Caseli, H.; Villavicencio, A.; Teixeira, A.; Perdigão, F. (Eds.) *Computational Processing of the Portuguese Language*, Proceedings of the 10th International Conference, PROPOR 2012, Coimbra, Portugal, April 17-20, 2012. http://www.propor2012.org/demos.html.

PAUMIER, S. 2003. De la reconnaissance des formes linguistiques à l'analyse syntaxique. PhD thesis, Université de Marne-la-Vallée, 2003.

PAUMIER, S. 2014. *Unitex 3.0 - User's Manual*. Paris: Université Paris-Est Marne-la-Vallée.

ROCHA, P. AND SANTOS, D. 2000. CETEMPúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa. In: Nunes, M.G. *et al.*, eds., *V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR'2000)*, São Paulo: ICMC/USP. pp. 131–140.

SAG, I.A., BALDWIN, T., BOND, F., COPESTAKE, A.; FLICKINGER, D. 2002. Multiword Expressions: A Pain in the Neck for NLP. In: Gelbukh, A. (ed.) *Proceedings of the Third International Conference, CICLing - Computational Linguistics and Intelligent Text Processing*, Mexico City, Mexico, February, 2002, pp. 1-15.