New developments on processing European Portuguese verbal idioms

Ana Galvão, Jorge Baptista, Nuno J Mamede

New developments on processing European Portuguese verbal idioms

Ana Galvão[1,3], **Jorge Baptista**[2,3], **Nuno J. Mamede**[1,3]

[1]Instituto Superior Técnico, Universidade de Lisboa
Av. Rovisco Pais 1, P-1049-001 Lisboa, Portugal

[2]Universidade do Algarve - Faculdade de Ciências Humanas e Sociais
Campus de Gambelas, P-2005-139 Faro, Portugal

[3]Instituto de Engenharia de Sistemas e Computação - INESC-ID Lisboa
L2F - Spoken Language Laboratory
R. Alves Redol 8, P-1000-029 Lisboa, Portugal

{a.s.galvao,Nuno.Mamede}@tecnico.ulisboa.pt, jbaptis@ualg.pt

Abstract. *This paper presents recent developments in processing verbal idioms within a rule-based grammar of European Portuguese. It describes the automatic construction of parsing rules directly from a lexicon-grammar matrix with about 2,500 idioms and about 100 structural, distributional, and transformational properties. Transformations (passive, pronominalization,* etc.*) of idioms' base sentences are now taken into account within the automatic rule generation process. An intrinsic evaluation achieves 95% recall.*

1. Introduction

Verbal idioms are a type of frozen sentences [Gross 1982, Gross 1996], where the verb and at least one of its arguments (subject or complement) are distributionally frozen together; and the global interpretation of the sentence is non-compositional, that is, it can not be calculated from the meaning the components of the idiom when they are used independently, e.g. *O Rui agarrou o touro pelos cornos* , lit.: 'Rui took/grasped the bull by the horns'. 'to take definite and determined action in order to deal with a difficult situation'. Parsing multiword expressions (MWE) such as verbal idioms is a challenging task for many Natural Language Processing (NLP) systems [Sag et al. 2002, Constant et al. 2017], since, for the most part, they have an internal structure identical to that of ordinary sentences, including one or more distributionally free arguments, and can undergo several, very general transformations, such as pronominalisation, passive, nominalisation, etc. Taking MWE into consideration, especially subsentential lexical units, can significantly improve the quality of several NLP tasks, like part-of-speech tagging [Constant and Sigogne 2011] or parsing [Constant et al. 2017]. Obviously, identifying MWE may lead to a more adequate representation of the meaning of a text. Most previous work deals with the identification of idioms and other MWE in texts [Ramisch et al. 2018, Ramisch et al. (eds.) 2018], since the low frequency of many verbal idioms in corpora makes spotting them a difficult task in lexicographic studies [Manning 1999, Pecina 2010]. The focus of this paper, however, will be on processing verbal idioms once they have already been integrated in a computational lexicon.

This paper presents new developments in the processing of European Portuguese (EP) verbal idioms, within the framework of a pipeline NLP system, STRING[1] [Mamede et al. 2012]. The paper's main contribution is the processing of the most common syntactic transformations (passive, pronominalization, *etc*.) accepted by these idioms. The paper is organized as follows: First, the lexicon-grammar of EP verbal idioms is presented, along with the parsing strategy adopted in STRING. .Next, the parsing of transformations is outlined. The paper, then, reports the results obtained in an *intrinsic* evaluation of the system, and concludes by pointing the challenges ahead.

2. Processing verbal idioms: current state

Previous work on European Portuguese verbal idioms [Baptista et al. 2004, Baptista et al. 2014, Baptista et al. 2016] done within Lexicon-Grammar framework [Gross 1982, Gross 1996], produced a lexicon-grammar matrix, currently with around 2,500 entries (one verbal idiom per line) along with the corresponding linguistic description. This description uses about 100 features (columns) to account for the structural, distributional and transformational properties of the idioms. The idioms are organized into 13 major classes (Table 1), according to the number of arguments selected by the verb in the frozen construction, and which arguments are distributionally free or frozen with the verb (the subject or one or more complements). For lack of space, the classification procedure is not provided in further detail here (see [Baptista et al. 2016, Galvão 2019]). The verb and the frozen elements of the idiom are explicitly encoded: the complements' preposition `Prep` (if any), the determiner `Det`, the frozen head noun `C`, and its left or right modifier `Modif`). The human/non-human nature of distributionally free arguments is represented by binary features ('+/−'), as well as several transformational properties. The transformations considered so far are all very general: passive, pronominalisation, symmetry, and dative restructuring (see below). Finally, all entries are illustrated by a manually produced example. These examples, while being perfectly natural and acceptable utterances, are 'artificial', almost 'laboratorial', since they contain all essential arguments (subject and complements) of the verb, and have been stripped of any spurious lexical material not relevant for the the interpretation of the idiom. Furthermore, whenever necessary, the verb is provided in an non-ambiguous inflected form, in order to prevent errors in previous stages of the processing, particularly in PoS tagging and disambiguation. These examples are also used for the intrinsic evaluation of the system.

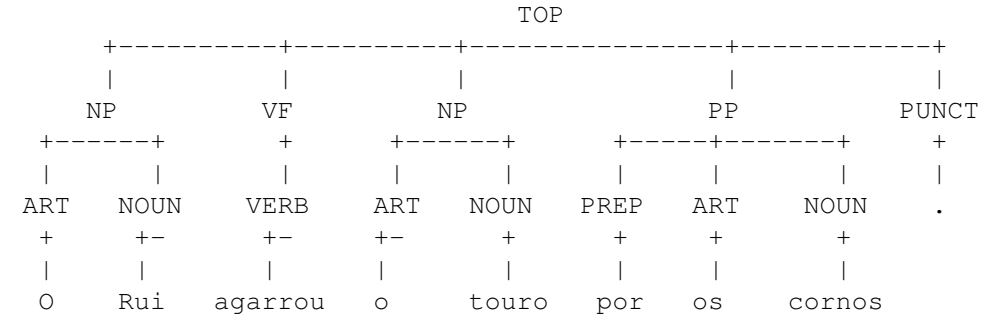
Table 1. Lexicon-Grammar of European Portuguese verbal idioms.

Class	Structure	Example	Translation/gloss	Count	%
C0	*C0 V w*	*O azar bateu à porta do Rui*	Bad luck knocked on Rui's door (have bad luck)	25	0,010
C1	*N0 V C1*	*O Rui bateu a bota*	Rui kicked the boot (died)	506	0,198
C1P2	*N0 V C1 Prep2 C2*	*O Rui comeu gato por lebre*	Rui ate cat for hare (was cheated)	284	0,111
C1PN	*N0 V C1 Prep2 N2*	*O Rui acertou agulhas com o Pedro*	Rui matched needles with Pedro (are in accord)	255	0,100
CADV	*N0 V ADV*	*O Rui vai longe*	Rui will go far (will be successful)	70	0,027
CAN	*N0 V (C de N)1 = C1 a N2*	*O Rui partiu os olhos de/a_a Ana*	Rui open the eyes of/to Ana (make understand)	182	0,071
CDN	*N0 V (C de N)1*	*O Rui veste a camisola da empresa*	Rui dones the t-shirt of the company (dedicate/loyal)	47	0,018
CNP2	*N0 V N1 Prep2 C2*	*O Rui conhece a Ana de nome*	Rui knows Ana by name (id)	175	0,068
CP1	*N0 V Prep1 C1*	*O Rui foi aos arames*	Rui went to the strings (be mad)	598	0,233
CPN	*N0 V Prep1 (N de C)1*	*O Rui foi na cantiga do Pedro*	Rui went in Pedro's song (be dupped)	103	0,040
CP1P2	*N0 V Prep1 C1 Prep2 C2*	*O Rui foi desta para melhor*	Rui went from this [one] to a better [one] (die)	170	0,066
CPP	*N0 V Prep1 C1 Prep2 N2*	*O Rui foi de Caifás para Pilatos*	Rui went from Caifas to Pilates (get in a worst situation)	77	0,030
CPPN	*N0 V C1 Prep1 C2 Prep3 C3*	*Isso deu água pela barba à Ana*	That gave water by the beard to Ana (very complicated)	51	0,020
CPPP	*N0 V Prep1 C1 Prep2 C2 Prep3 C3*	*O Rui contava com o ovo no cu da galinha*	Rui counted with the egg in the ckiken's ass (be too confident)	5	0,002
CV	*N0 V Vinf w*	*Esta rua vai dar à praça*	This street goes give to the square (lead to)	13	0,005
			Total	2,562	

[1] `https://string.l2f.inesc-id.pt/` (last access: 06/08/2019)

Since the construction of the lexicon-grammar matrix is not only a complex process but it is also carried out manually, it is thus a very error-prone task. To reduce the human error in this process, an *Automatic Validator* has been built, written in Perl, to check the formal consistency of the matrix. This validator takes as input the CSV-converted lexicon-grammar matrix and performs the following checks, outputing the corresponding error messages: (i) *cell* content validation: checks if the content of the cell in a given column is consistent with the predefined values for that column; (ii) *class* consistency cross-validation: depending on the class of the idiom, the number of relevant columns and the values therein can vary; and (iii) related properties cross-validation: consistency among related properties, represented in different columns, is checked. The validator resorts to a set of several dozens of manually crafted rules. Based on the error messages outputted by the validator, it is possible to detect most input errors, which are then manually corrected. When no formal errors are found, the matrix is ready to be processed.

The processing of verbal idioms is done within the framework of the rule-based, Xerox Incremental Parser (XIP) [Ait-Mokhtar et al. 2002], which is the parsing module of the NLP system pipeline STRING [Mamede et al. 2012], developed for Portuguese. This system performs all the basic text processing tasks: (i) text segmentation and tokenisation; (ii) part-of-speech (PoS) tagging; (iii) rule-based and statistical PoS disambiguation; and (iv) parsing. The later includes both *chunking* and dependency parsing. The first forms the elementary constituents (e.g. *chunks*: noun phrase, `NP`; prepositional phrase, `PP`; *etc*.). The second extracts the relations between the chunks' heads, e.g. subject (`SUBJ`). STRING also performs other, common NLP tasks, such as named entity recognition (identification and classification, *NER*), anaphora resolution, event detection, among others. Fig. 1 illustrates the parse tree of the example above, with the chunks and some dependencies calculated by the parser.




```
MAIN(agarrou)                    MOD_POST(agarrou,cornos)
DETD(Rui,O)                      SUBJ_PRE(agarrou,Rui)
DETD(touro,o)                    CDIR_POST(agarrou,touro)
DETD(cornos,os)                  FIXED(agarrou,touro,cornos)
VDOMAIN(agarrou,agarrou)         NE_PEOPLE_INDIVIDUAL(Rui)

0>TOP{NP{O Rui} VF{agarrou} NP{o touro} PP{por os cornos} .}
```

Figure 1. Parse tree of a verbal idiom.

Figure 1 shows some auxiliary dependencies, such as the determiner `DETD`, linking the articles to the head nouns of the `NP` and `PP` chunks. There is also a `PREPD` dependency (not shown) linking the preposition *por* 'by' to the head noun of the `PP`. The

VDOMAIN dependency links the first verb of a chain of auxiliary verbs to the main verb [Baptista et al. 2010]. The main dependencies, *v.g.* subject (SUBJ), modifier (MOD), and direct complement (CDIR), are also calculated, linking the verb to its arguments. The named entity ***Rui*** is also captured by the unary dependency NE, which takes the features _PEOPLE and _INDIVIDUAL corresponding to the entity type [Hagège et al. 2008].

Since, for the most part, verbal idioms comply with the general rules governing the structure of well-formed sentences in the language, including word order, it has been deemed more appropriate [Rassi et al. 2014, Baptista et al. 2014] to detect them only at a later stage of the parsing process, after the main syntactic dependencies between the sentences's constituents have been calculated. These dependencies are then used to detect idioms. If an idiom is detected, the system produces a FIXED dependency, whose arguments are the the main verb and its frozen arguments. The following rule, shown in Fig. 2), was fired and captured the idiom of this example:

```
if (  VDOMAIN(#?,#2[lemma:agarrar])  &
      CDIR[post](#2,#3[surface:touro])  &
              DETD(#3,?[surface:o])  &
      MOD[post](#2,#4[surface:cornos])  &
              PREPD(#4,?[surface:por])  &
              DETD(#4,?[surface:os])
    )
    FIXED(#2, #3, #4)
```

Figure 2. Parsing rule for the verbal idiom *agarrar o touro pelos cornos* 'take/grab the bull by the horns'.

This parsing rule has two parts: the first is a if() structure of conditions that must be satisfied so that the consequent of the rule is triggered; the consequent writes the FIXED dependency and its arguments. The VDOMAIN is used to capture the main verb, even if this is construed with a chain of auxiliary verbs. The next dependencies verify if the heads of direct complement CDIR and the prepositional phrase modifier MOD are the same as encoded in the lexicon-grammar matrix. The same applies to the auxiliary dependencies DETD within the complements and the PREPD dependency for the MOD. All conditions are joined by '&' (disjuction '||'may also be used) The variables (signaled by '#') are then used to produce the FIXED dependency.

The process of automatically generating the parsing rules directly form the linguistic information encoded in the matrix is quite complex, so it will only be sketched here. First, each column of the matriz is associated with a XIP rule. Each relevant column value contributes with a condition to that rule (relevant columns depend on the verbal idiom class); and each main constituent's head is associated to a variable (by convention, the subject is associated with variable #1, the verb to #2, and so forth). The system systematically explores the properties encoded in all columns of the matrix, adding the corresponding conditions to the if() structure of the parsing rule. Finally, it writes the FIXED dependency with its arguments. The output of this module consists of the parsing rules, the corresponding manually produced example, and the expected output of the FIXED dependency.

Already at a first attempt to integrate verbal idioms [Baptista et al. 2016] into the STRING, a special module had been built to automatically generate the parsing rules for

the detection of idioms, based on the information directly extracted from the lexicon-grammar matrix. However, the method proved to be very rigid, as it depended on the column number and order. Also, only the passive transformation was considered. Furthermore, the evaluation was carried out sentence by sentence, and each time the system had to be initialised. This took too much time to be practical. Besides, this evaluation only reported whether the `FIXED` dependency had been extracted or not, thus providing limited feedback for the further development of STRING. The new developments, presented below, address all these issues.

3. Processing verbal idioms: new developments

Several improvements were introduced in the processing of idioms within STRING. These new developments are presented next. Foremost, the system is now able to process several sentence transformations: Different types of *pronominalisation* of the distributionally free complements are considered, named after the case/type of the pronoun involved: (i) *accusative* `[PronA]`: *O Pedro lançou* ***o João*** *às feras* (class CNP2) 'Pedro threw João to the wolves' = *O Pedro lançou****-o*** *às feras* 'Pedro threw him to the wolves'; (ii) *dative* `[PronR]`: *O Pedro lançou a escada* ***à Ana*** (CNP2) lit.: 'Pedro threw the stairs to Ana' 'Pedro tried to seduce Ana' = *O Pedro lançou****-lhe*** *a escada* lit.: 'Pedro threw to-her the stairs'; (iii) *reflexive* `[PronR]`: *O Pedro reduziu* ***o João*** *ao silêncio* (CNP2) lit.: 'Pedro reduced João to silence' cp. *O Pedro reduziu****-se*** *ao silêncio* lit.: 'Pedro reduced himself to silence'; (iv) *possessive* `[PronPos]`: *O Pedro abriu os horizontes da Ana* (CAN) lit: 'Pedro opened the horizons of Ana' = *O Pedro abriu os seus horizontes* 'Pedro open her horizons'. Two types of passive sentences, with different auxiliaries: (v) *Passive* with auxiliary *ser* 'be'; in this type of passive, the subject of the active sentence becomes a prepositional complement *por N* 'by N'; only the verb *ser* 'be' is considered in this case: *O João foi reduzido ao silêncio pelo Pedro* (CNP2) 'João was reduced to silence by Pedro'; (vi) *Passive* with auxiliary *estar* 'be'; in this passive, the subject of the active sentence is usually zeroed; any other copula verb, including *ficar* 'become', can also be captured in this rule: *O João estava/ficou reduzido ao silêncio* (CNP2) 'João was reduced to silence'. And, finally, the dative restructuring: (vii) *dative restructuring* `[Rdat]` [Leclère 1995]: this type of transformation splits a complex complement *(*N_a *de* N_b*)*$_1$ 'N of N' into two constituents, *(*N_a*)*$_1$ *(a* N_b*)*$_2$ 'N to N', the noun's complement becoming a dative (indirect) complement, more closely attached to the verb: *O Pedro abriu os horizontes* ***da Ana*** (CAN) lit: 'Pedro opened the horizons of Ana' = *O Pedro abriu os horizontes* ***à Ana*** 'Pedro open the horizons to Ana', which can now undergo the dative pronominalisation: = *O Pedro abriu****-lhe*** *os horizontes* 'Pedro open her the horizons'. Symmetric constructions [Borillo 1971, Baptista 2005] involve the coordination of two constituents, e.g. *O Rui juntou os trapinhos com a Ana* (C1PN), lit.:'Rui got his rags together with Ana' = *O Rui e a Ana juntaram os trapinhos*, lit.:'Rui and Ana got his rags together ', 'Rui and Ana got married/together'. They were described by manually crafted rules , not only for the small number of symmetric verbal idioms found so far, but also for the complexity involved in capturing the coordinated arguments and the (facultative) presence of an echo complement, *um Prep outro* [Baptista and Mamede 2013].

When a verbal idiom accepts one of these transformations, the rule generator produces a disjunction '`||`' in the `if()` structure. For example, the first example *O Pedro lançou o João às feras* 'Pedro threw João to the wolves' accepts both the accusative and the reflexive pronominalization of the direct complement `CDIR`, so this line becomes:

```
( CDIR[post](#2,#3[UMB-Human]) ) || CLITIC(#2,?[ref]) || CLITIC(#2,#3[acc]) )  &...
```

For passive transformations, a new rule is produced because of the changes in the set of dependencies and their arguments that such structurally different sentences entail.

A *configuration file* enables the user to define which restrictions are to be applied to generate the transformation rules. Controllable restrictions apply to determinants, prepositions and both left and right modifiers of the frozen head noun; the distributional constraints to any of the free constituents, both the subject and/or the complements, can also be taken into account or ignored. In the barest configuration, only the major dependencies between the verb and the head nouns of the frozen constituents are included in the rules.

Secondly, a rule-based *Automatic Example Generator* was build from scratch, which produces a simple example for each transformation that can be applied to a given idiom, based on the linguistic information encoded in the matrix. These 'artificial' examples allow the linguist to better perceive the adequacy of his/her (theoretical) description, and are also used to evaluate the system. Sentences are generated along with the correct `FIXED` dependency, used for reference. For a better perception of the system's performance, the sentences produced for each transformation are kept apart. The examples produced by this rule-based generator were manually checked by a linguist, who signalled the grammatical errors (e.g. missing contractions or prepositions) or inconsistencies produced (missing pronouns, complements); the code was, then, revised to correct those errors and a new set of sentences was generated, in a iterative way, until a 'cleaner' output was generated. In this process, several inconsistencies in the linguistic data could also be resolved. In all, 1,170 transformationally-derived sentences were automatically generated. For lack of space, the break down of transformations per class can not be provided here. Please refer to [Galvão 2019] for further details.

Finally, an *Evaluation Module* was build anew, which performs an intrinsic evaluation of the system. It also now takes into consideration 3 levels of granularity in the results: (i) whether the `FIXED` dependency was captured or not; (ii) if the number of its arguments is correct (`NB-ARG`); and (iii) if the arguments are the same as those in the reference (`ARG`). To this end, this module takes as input the two sets of examples, those manually produced along the verbal idioms' entries; and those automatically generated for the transformations. The sentences are processed through STRING in a single batch, and the output is then compared with the reference.

In order to achieve a more comprehensive evaluation of the system, 20% of the base sentences that constitute the examples in the matrix were randomly selected from each class of idioms, totalling 511 sentences, and they were then subject to different types of modifications. These, manually produced, modifications aimed at creating incorrect/unacceptable ('*') or non-idiomatic ('º'), but still similar, examples in order to test whether STRING still incorrectly extracts the `FIXED` feature in spite of them. Examples of these changes are, for the idiom *O Rui agarrou o touro pelos cornos* 'Rui grabed the bull thy the horns' (C1P2): the human/non-human nature of the different distributionally free arguments; changing or zeroing the preposition (º/**O Rui agarrou* ***n****o touro* ***a****os cornos*), the determinant or the modifier of the frozen head noun of an argument; removing one or more frozen constituents (º*O Rui agarrou o touro*); changing the case of the pronominalized constituent (in a way unacceptable by the idiom); etc.

4. Results

Table 2. Results for verbal idioms identification: Manually produced sentences.

Class	Total	#FIXED	%	#NB-ARG	%	#ARG	%
CADV	16	7	43.8	7	43.8	5	31.3
C0	21	15	71.4	15	71.4	12	57.1
C1	503	484	96.2	481	95.6	448	89.1
CAN	182	156	85.7	156	85.7	153	84.1
CDN	46	37	80.4	37	80.4	36	78.3
C1P2	291	274	94.2	266	91.4	228	78.4
C1PN	259	224	86.5	216	83.4	206	79.5
CNP2	176	152	86.4	151	85.8	149	84.7
CP1	718	635	88.4	628	87.5	558	77.7
CPN	106	74	69.8	71	67.0	63	59.4
CPP	195	130	66.7	126	64.6	115	59.0
CPPN	36	28	77.8	27	75.0	26	72.2
CV	12	6	50.0	4	33.3	4	33.3
Total	2,561	2,222	86.8	2,185	85.3	2,003	78.2

Table 2 shows the STRING results from the manually produced sentences in a first run. Overall, recall varies from 86.8% with the more relaxed criterion of just capturing the FIXED dependency; to 85.3, when the number or the dependency's arguments (NB−ARG) is considered; down to 78.2% for a complete match of the dependency's arguments (ARG). Many of these errors are due to previous processing steps in the pipeline, especially POS-tagging and disambiguation.

Table 3. Results for verbal idioms identification: Automatically generated, transformation-derived sentences.

Transformation	Total	#FIXED	%	#NB-ARG	%	#ARG	%
PronA	187	170	90.9	169	90.4	165	88.2
PronD	178	131	73.6	130	73.0	129	72.5
PronPos	324	268	82.7	266	82.1	265	81.8
Rdat	192	107	55.7	106	55.2	106	55.2
PassSer	185	142	76.8	141	76.2	139	75.1
PassEstar	83	70	84.3	69	83.1	68	81.9
Total	1,170	909	77.7	902	77.1	884	75.6

Table 3 shows the results of the evaluation on the set of transformation-derived, automatically generated sentences. Notice that only the idiom accepting each transformation were considered for each class, so that the total number of idioms per class varies depending on the transformation being evaluated. Global results, even if somewhat inferior (77.7% for FIXED), are similar to those found for the manually produced sentences, especially in the strictest criterion (75.6%).

For a second run, several duplicate entries of the matrix were either removed or corrected. Some of these duplicates resulted from indicating the lemma instead of the surface form of a given frozen element. Obvious input errors, like the verb in the matrix being different from the verb in the example, were also correct. The transformation-derived sentences were automatically generated again and integrated in this run. Results, shown in Table

4, indicate an overall 95.1% recall when only the FIXED dependency is considered, and 92.5% when there is a perfect match, including the arguments of the dependency.

Table 4. Results for the 2^{nd} run evaluation.

Sentences	Count	FIXED	%	NB-ARG	%	ARG	%
base	2,542	2,429	0.956	2,400	0.944	2,337	0.919
transformed	1,157	1,088	0.940	1,083	0.936	1,083	0.936
Total	3,699	3,517	0.951	3,483	0.942	3,420	0.925

Finally, the system was also run over the 511 modified based sentences, with the system configured to the consider only the verb and frozen constituents' head nouns, and to ignore the distributional properties of the free arguments, the determiners and modifiers of the frozen head nouns. The FIXED dependency was not extracted from 481 sentences (94.1%). However, in spite of the changes introduced, for 30 sentences, the system still extracts the FIXED dependency incorrectly. Some of these wrong cases are due, as expected, to the settings defined in the configuration file, e.g. determiners being ignored, as in *O Rui engoliu **esses** sapos* lit: 'Rui swallowed those frogs' 'eat crow'. Also, the absence of a obligatory complement, as in *O João fechou* [*algo*] *a sete chaves* 'João closed [something] with seven keys' 'safely locked/under lock and key' is not enough to preclude the extraction of the FIXED dependency. In the future, we intend to produce a new example generator module, to systematically explore all the variations considered in this small sample file and test it against the lexicon-grammar using all the XIP rule-generator configurations envisaged.

5. Conclusion and future work

This paper presented the new developments in the parsing of verbal idioms in European Portuguese. Several improvements were introduced in the STRING processing chain, namely in the automatic parsing rules' generator, which is now able to produce rules that capture several transformation-derived sentences. Also, an automatic example generator was produced anew, which builds these transformed sentences from the information encoded in the lexicon-grammar matrix. A new automatic evaluator was built, featuring a more granular assessment of the system's performance, including not only the extraction of the FIXED dependency, but also its correct number of arguments, and the correspondence to the arguments stated in the reference. At a first run, results were already very promising, reaching 78.2% recall in the strictest evaluation (exact match) and 86.8% in the relaxed mode (only the dependency). Similar, though lower results were obtained for the transformed sentences: 75.6% recall (for exact match) and 77.7% (in the relaxed mode). After error analysis and correction, it was possible to improve these results, both for the base and the transformed sentences. A second run of the system produced 92.5% recall in the exact match scenario, and 95.1% in the relaxed mode. In the near future, a similar procedure to integrate the lexicon-grammar of verbal idioms from Brazilian Portuguese [Vale 2001], as well as an *extrinsic* evaluation of the system are being envisaged, using the data sets of [Baptista et al. 2014] and [Ramisch et al. (eds.) 2018, Ramisch et al. 2018].

Aknowledgments

Research for this paper was partially supported with public funds by Fundação para a Ciência e a Tecnologia (FCT) through program ref. UID/CEC/50021/2019.

References

Ait-Mokhtar, S., Chanod, J., and Roux, C. (2002). Robustness beyond shallowness: incremental dependency parsing. *Natural Language Engineering*, 8(2/3):121–144.

Baptista, J. (2005). Construções simétricas: argumentos e complementos. In *Estudos de Homenagem a Mário Vilela*, pages 353–367. Campo das Letras, Porto.

Baptista, J., Correia, A., and Fernandes, G. (2004). Frozen sentences of portuguese: Formal descriptions for NLP. In *Workshop on Multiword Expressions: Integrating Processing*, pages 72–79. ACL.

Baptista, J., Fernandes, G., Talhadas, R., Dias, F., and Mamede, N. (2016). Implementing European Portuguese Verbal Idioms in a Natural Language Processing System. In Corpas Pastor, G., editor, *Computerised and Corpus-based Approaches to Phraseology: Monolingual and Multilingual Perspectives*, pages 102–115. Proceedings of EUROPHRAS 2015.

Baptista, J. and Mamede, N. (2013). Reciprocal Echo Complements in Portuguese: Linguistic Description in view of Rule-based Parsing. In Baptista, J. and Monteleone, M., editors, *Proceedings of the 32nd International Conference on Lexis and Grammar (CLG'2013)*, pages 33–40, Faro, Portugal. CLG'2103, Universidade do Algarve – FCHS.

Baptista, J., Mamede, N., and Gomes, F. (2010). Auxiliary verbs and verbal chains in European Portuguese. In *Computational Processing of the Portuguese Language (PROPOR 2010)*, number 6001 in LNAI/LNCS, pages 110–119.

Baptista, J., Mamede, N., and Markov, I. (2014). Integrating verbal idioms into an NLP system. In *Computational Processing of the Portuguese Language (PROPOR 2014)*, volume 8775 of *LNAI/LNCS*, pages 251–256.

Borillo, A. (1971). Remarques sur les verbes symétriques. *Langue Française*, 11(1):17–31.

Constant, M., Eryigit, G., Monti, J., van der Plas, L., Ramisch, C., Rosner, M., and Todirascu, A. (2017). Multiword expression processing: A survey. *Computational Linguistics*, (837–892).

Constant, M. and Sigogne, A. (2011). Mwu-aware part-of-speech tagging with a crf model and lexical resources. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, pages 49–56, Portland, Oregon, USA. Association for Computational Linguistics.

Galvão, A. (2019). Processar expressões fixas em português: Geração automática de regras e exemplos a partir de um léxico-gramática. Master's thesis, Universidade de Lisboa – Instituto Superior Técnico, Lisboa.

Gross, M. (1982). Une classification des phrases ≪figées≫ du français. *Revue Québécoise de Linguistique*, 11-2:151–185.

Gross, M. (1996). Lexicon-grammar. In Brown, K. and Miller, J., editors, *Concise Encyclopedia of Syntactic Theories*, pages 244–259. Pergamon, Cambridge.

Hagège, C., Baptista, J., and Mamede, N. J. (2008). Reconhecimento de entidades mencionadas com o XIP: Uma colaboração entre o INESC-L2F e a Xerox. In Mota, C. and

Santos, D., editors, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, pages 261–274. Linguateca.

Leclère, C. (1995). Sur une restructuration dative. *Language Research*, (31-1):179–198.

Mamede, N., Baptista, J., Diniz, C., and Cabarrão, V. (2012). STRING - A Hybrid Statistical and Rule-Based Natural Language Processing Chain for Portuguese. In Abad, A., editor, *International Conference on Computational Processing of Portuguese (PROPOR 2012) - Demo Session*, Coimbra, Portugal. http://www.propor2012.org/demos/DemoSTRING.pdf.

Manning, Chris; Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, 1st edition.

Pecina, P. (2010). Lexical association measures and collocation extraction. *Language Resources and Evaluation*, 44:137–158.

Ramisch, C., Ramisch, R., Zilio, L., Villavicencio, A., and Cordeiro, S. (2018). A Corpus Study of Verbal Multiword Expressions in Brazilian Portuguese. In *Computational Processing of the Portuguese Language (PROPOR 2018)*, volume 11122 of *LNAI/LNCS*, pages 24—34.

Ramisch et al. (eds.), C. (2018). Edition 1.1 of the PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressionshared Task on Automatic Identification of Verbal Multiword Expressions. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 222–240, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Rassi, A., Santos-Turati, C., Baptista, J., Mamede, N., and Vale, O. (2014). The fuzzy boundaries of operator verb and support verb constructions with *dar* "give" and *ter* "have" in Brazilian Portuguese. In *Proceedings of Workshop on Lexical and Grammatical Resources for Language Processing, COLING 2014*, pages 92–101. ACL.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. (2002). Multiword Expressions: A Pain in the Neck for NLP. In *Proceedings of Computational Linguistics and Intelligent Text Processing*, volume 2276 of *LNAI/LNCS*, pages 1–15, Berlin. 3rd International Conference CICLing-2002, Springer.

Vale, O. A. (2001). *Expressões Cristalizadas do Português do Brasil: uma proposta de tipologia*. Tese de Doutorado, Universidade Estadual Paulista, Araraquara.