PORTUGUESE PROVERBS: TYPES AND VARIANTS

Sónia Reis
University of Algarve
reis.soniamm @gmail.com

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

Keywords: Proverbs, European Portuguese, Variation, Automatic Identification, Corpus Linguistics

Abstract

Drawing on the methodology and previous results of Rassi *et al.* (2014) on the automatic identification of Brazilian Portuguese proverbs, this paper reports on an extension of that experiment, but now focused on the identification of the European Portuguese proverbs and their variants. Based on a large collection of over 56 thousand Portuguese proverbs and their variants, a database of proverb types was specifically built for natural language processing, along with the finite-state tools that allow for the identification of these strings in texts. Our aim is to make these linguistic resources and language processing tools publicly available, which will undoubtedly be deemed useful assets to other paremiologic studies.

1. INTRODUCTION

Proverbs are an important part of most societies' culture and language. As micro-texts, brought into discourse from the common cultural repository, they are subject to many creative types of variation. On the other hand, functioning as in quotation mode, they integrate discourse in an almost disruptive way, challenging natural language processing (NLP) systems, and requiring their accurate identification and delimitation.

Concerning Portuguese proverbs, and though several, extensive collections of proverbs are available in printed form (Machado, 2011), to the best of our knowledge, no resources have been specifically produced for NLP purposes, even if some digitally available dictionaries (Almeida, 2014) include a few examples, interspersed between other type expressions, like different types of idioms and many forms of slang.

Recently, Rassi *et al.* (2014) have proposed a formal (syntactic) classification of proverbs, based on a collection of over 3,500 proverb variants, organized in 594 proverb types (Rassi, 2014), and taken from several dictionaries from the Brazilian variety of the Portuguese language. The authors presented a finite-state based method for the automatic identification of proverbs in large-sized corpora, and experimented on a 29M tokens corpus of journalistic text (Bruckschein *et al.* 2008), taken from the daily online edition of the Brazilian newspaper *Folha de São Paulo*. The authors report a 60 to 73% precision, depending on the proverb class and the width of the insertion window between the proverbs' keywords. In spite of the corpus size, but not surprisingly, only 137 types and

788 instances were matched, most likely because of the journalistic nature of the texts in this corpus. However, seen from this side of the Atlantic, results from Rassi and colleagues are surprising mostly for the fact that, in spite of some American idiosyncrasies, most proverbs seem to exist also in the European variety, quite unlike the mismatch that has been found for verbal idioms (undisclosed reference).

Drawing on this methodology and these previous results, this paper reports on an extension of that experiment, but now focused on the identification of the European Portuguese proverbs and their variants. This intends to set up the basis for a large collection of Portuguese proverbs and their variants, specifically built for natural language processing, and to make it publicly available, along with the finite-state tools built for retrieving them from texts. These tools and resources will undoubtedly be deemed useful assets to other paremiology studies.

The remainder of this paper is structured as follows: Section 2 presents the methods, starting with the formal classification criteria, and the current state of the collection of Portuguese proverbs' types (§2.1); followed by the criteria to select the proverbs' keywords in order to define those proverbs' types (§2.2); next, the finite-state tools to match proverbs in text are presented (§2.3), followed by the proverbs' collections used to produce a digitized list of Portuguese proverbs and variants (§2.4). Section 3 present some preliminary evaluation of the finite-state tools when applied to this list and discusses some of the issues pertaining to the resulting matches. Finally, Section 4 concludes the paper and hints at future work.

2. METHODS

2.1 Adapting the proverbs classification of Rassi *et al.* (2014) to European Portuguese

Based on the syntactic classification of Brazilian Portuguese proverbs proposed by Rassi *et al.* (2014) we produced a database, in tabular format, with the key elements of each proverb. This formal, taxonomic, approach allows us to determine accurately the concept of variant and base form of a proverb.

In order to better frame this paper, we first present the formal/syntactic classification proposed by Rassi *et al.* (2014) of Brazilian Portuguese proverbs, which may also be adopted by and large to European Portuguese proverbs.

This classification is based on the number of propositions forming the proverb and the part-of-speech (PoS) of their main elements. Other, secondary features are also used. The number of propositions (or clauses) organizes the data in 3 main classes (**P1x**, **P2x** and **P3**). The specific sentence type of **P1x**, or the transformations it may (or may not) undergo are used to further split this type into several classes. Thus, we find the following classes:

P1F1: impersonal constructions; while in European Portuguese this are mostly sentences with the verb *haver* (there be), in Brazilian Portuguese one also finds an impersonal use of *ter* (have); the head noun is often modified by a prepositional phrase (PP) or even a relative subclause; impersonal constructions with indefinite clitic pronoun *se* were also included in this class:

Não há rosa sem espinhos 'There is no rose without thorns'

Não há mal que sempre dure 'There is no evil that lasts forever';

Tem muita estrela pra pouco céu 'There is to many star[s] for little/small sky'

Devagar se vai ao longe 'Slowly one can go far'

P1F2: attributive sentences with copula verb *ser*; the subject is usually a noun (sometimes a verb in the impersonal infinitive); the predicative element can be a noun, an adjective, an verb in the impersonal infinitive, or even a prepositional phrase:

A fome é o melhor tempero 'Hunger is the best seasoning'

O amor é cego 'Love is blind'

Partir é morrer um pouco 'To leave is to die a little'

O silêncio é de ouro 'Silence is of gold'

P1F3: direct-transitive verb constructions:

Muitos cozinheiros estragam a sopa '[Too] many cooks ruin the soup'

Os fins justificam os meios 'The ends justify the means'

Uma mão lava a outra 'One hand washes the other'

P1F4: obligatory negation:

Burro velho não aprende línguas 'Old donkey does not learn languages'

Uma andorinha não faz a primavera 'A single swallow does not make spring'

Gostos não se discutem 'Tastes are not discussed'

P1F5: obligatory fronting of the prepositional complement:

De boas intenções está o inferno cheio 'Hell is full of good intentions'

Em terra de cegos quem tem um olho é rei 'In a land of blind [people], he who has an eye/the one-eyed is king"

Em boca fechada não entra mosca 'In [a] closed mouth no fly enters'

Next, two-clause proverbs are considered (**P2x** classes). These include:

P2F1: proverbs with a main clause and a comparative subordinate clause; as comparison can be expressed in many different ways, different types of comparative structures are considered:

Antes tarde (do) que nunca 'Better late than never'

Mais vale um pássaro na mão do que dois a voar 'Better a bird in the hand that two fling"

Não há pior cego do que aquele que não quer ver 'There is no one more blind that the one who does not want to see'

P2F2: proverbs with two coordinate clauses; in some cases, the coordinative conjunction is not expressed but only implied (asyndeton):

As moscas mudam mas a merda é sempre a mesma 'The flies change but the sheet is always the same'

Vão-se os aneis [*mas, e*] *ficam os dedos* 'The rings go away (but/and) the fingers stay'

Deus não fecha uma porta que não abra logo duas 'God does not close a door without opening two right away'

P2F3: verb-less proverbs with two phrases; in some cases, a coordination or subordination nexus can be inferred but the coordinate/subordinate conjunction is not expressed,

only implied (asyndeton); in other cases, the (implied) main verb of the sentence can be inferred:

Cada roca com seu fuso, cada terra com seu uso 'Each spinning distaff with its spindle, each land with its ways

Muito riso, pouco siso '[Too] much laughter too little judgement'

Tal pai, tal filho 'Like father, like son'

Olho por olho, dente por dente 'Eye for eye, tooth for a tooth'

Cada cabeça [?dá] *sua sentença* 'Each head [gives] its sentence (opinion)'

P2F4: pseudo-interrogative sub-clauses, introduced by interrogative *Qu-* (*Wh-*) pro-forms:

O que não mata engorda 'What doesn't kill [one] makes [one] fat'

Quem avisa amigo é 'He who warns [one] is a friend'

P2F5: proverbs with a main clause and a subordinate clause:

Fazer o bem sem olhar a quem 'To do good without looking to whom'

Devagar que tenho pressa 'Slowly for I am in a hurry'

P2F6: proverbs with a main clause and an obligatorily fronted subordinate clause:

Para morrer basta estar vivo 'In order to die, is enough to be alive'

Enquanto houver vida há esperança 'While there is life, there is hope'

Quando a esmola é muita, o pobre desconfia 'When the charity is too much, the poor gets suspicious'

Finally, class **P3** includes the long proverbs with more than 2 clauses/propositions. Unlike the previous classes, this one has not been subdivided yet, pending on a accumulation of data that would render such sub-classification necessary. Thus, in **P3** we find:

P2F6: proverbs with more than 2 clauses/propositions; the coordinative conjunction can be omitted; often, instead of a clause one finds verb-less phrases:

Mãos frias, coração quente, amor ardente 'Cold hands, hot heart, burning love'

Laranja, de manhã é ouro, à tarde é prata e à noite mata 'Orange, in the morning is gold, in the afternoon is silver and at night it kills'

Um é pouco, dois é bom, três é demais 'One is [too] little, two is good, three is too much'

Following their publication, A. Rassi and her co-authors published the list of 594 proverb types and their classification[44], which we used as the basis for our own classification of European Portuguese proverbs. While, in its general features, their taxonomy seems a useful tool to organize the complex and abundant data already available, in some cases, we disagreed with the authors' classification and decided to assign some of their proverbs to a different class. We also some added Portuguese-specific proverbs and variants, absent from Rassi et al. (2014) list. This careful and close review of the proverbs list, along with the growing number of classified proverbs, will eventually lead to a more granular classification, especially for the larger classes, that is, those with the larger number of types.

[44] https://www.researchgate.net/publication/266852580_Rassi2014?ev=prf_pub.

For instance, the proverbs with the copula verb *ser* (be) constitute such a large set that it could be advantageous to create sub-classes according to the morphological class (PoS) of the element that are associated (adjective, noun, etc.). Table 1 shows the classification in its current state:

Class	Structure	definition	Example/gloss (or wbw translation)	types	%
P1F1	0 V w	impersonal constructions	*Não há rosa sem espinhos.* 'There is no rose without thorns'	**20**	0.03
P1F2	N_0 Vcop (N+Adj)	predicative constr. (copula verb)	*O amor é cego.* 'Love is blind'	**53**	0.09
P1F3	N_0 V N_1	direct transitive (no PP)	*Uma mão lava a outra.* 'One hand washes the other'	**80**	0.13
P1F4	N_0 Neg V w	obligatory negation	*Uma andorinha não faz a primavera.* 'a single sparrow does not makes spring'	**53**	0.09
P1F5	Prep N_1, N_0 V	obligatory fronting of PP	*Pela boca morre o peixe.* 'By the mouth dies the fish'	**45**	0.08
P2F1	F_1 Cs-comp F_2	comparative	*Mais vale um pássaro na mão do que dois a voar.* 'Better a bird in the hand than two flying'	**39**	0.07
P2F2	F_1 (Cc) F_2	coordinate (asyndeton)	Vão-se os aneis, ficam-se os dedos. 'The rings go away, the fingers stay'	**71**	0.12
P2F3	N_1 (Cc) N_2	coordinate (w/o verb)	*Tal pai, tal filho.* 'Like father, like son'	**48**	0.08
P2F4	*Quem V V w*	interrogative subject *Qu-* 'Wh-'	*Quem vê caras não vê corações.* 'Who sees faces does not see hearts'	**90**	0.15
P2F5	F_1 Cs F_2	subordinate	*Fazer o bem sem olhar a quem.* 'To do good without looking to whom'	**20**	0.03
P2F6	Cs F_2, F_1	obligatory fronting of subordinate	*Quando um burro fala, os outros abaixam as orelhas.* 'When a dunkey speaks, the others lower their ears'	**28**	0.05
P3	F_1 C F_2 C F_3	3-clause	*Mãos frias, coração quente, amor ardente.* 'Cold hands, warm heart, burning love'	47	0,08
				594	

Table 1. Classification of Portuguese proverbs (adapted from Rassi *et al.* 2014).

2.2. Compiling the proverbs' keywords

For each class, we defined the key elements that were to be matched in order to unambiguously identify the proverb. As Rassi and co-authors had not published this data (only an example per type and its class are provided), we have re-done most of their work, carefully reviewing the selection criteria that help define the proverb's core elements. These keywords vary depending on the class and, in some cases, even on the proverb itself and its variants. Therefore, only a glimpse of the complex process of selecting the keywords can be provided here.

For keywords, the main content words (nouns, verbs and adjectives) are usually selected, and represented by their lemma (represented inside chevrons, '<' and '>'), rather than the surface (inflected) form, in order to allow for the capture of creative reuse of the proverb. Hence in, for the entry

[**P1F3**] *Os* ***fins*** *(não)* ***justificam*** *os* ***meios*** 'The ends (do not) justify the means'

the lemmas of the two nouns, *fim* and *meio*, and the verb are considered keywords:

<fim> <justificar> <meio>

Usually, copula verbs and auxiliaries are dropped, as in:

[**P3**] ***Um*** *é* ***pouco, dois*** *é* ***bom, três*** *é* ***demais***

'One is too few, two is good, three is too much'

where only the subjects and the adjectives are kept; notice that in this case, the numerals are not determining an noun, as determinants are usually discarded; the specific word order is characteristic of the proverbial nature of the sentence:

<um> <pouco> <dois> <bom> <três> <demais>

In some cases, it is the structure, rather than the specific words, that is key to the proverb. For example, in the proverb:

[**P2F1**] ***Duas cabeças pensam melhor do que uma*** 'two heads think better than one'

we can group the many variants using the following string[45]:

<dois> <N:p> <V:p> (mais+melhor) (do+<E>) que <um:s>

where <N:p> and <V:p> stand for any noun and verb (in the plural), the comparative adverbs *mais* '*more'* and *melhor* 'better' introduce the subordinate comparative conjunction (*do*) *que* 'than', which allows for the zeroing of *do*, and the numeral *um* 'one'.

The number of variants of a proverb can yield quite complex expressions, as in the next case (all variants taken from proverbs' collections; see §2.3, below):

[**P1F4**] ***Não fales*** *de* ***corda*** *em* ***casa*** *de* ***enforcado***

Não *se deve* ***falar*** *em* ***corda*** *em* ***casa*** *de* ***enforcado***

Não *se* ***fala*** *em* ***corda*** *em* ***casa*** *de* ***enforcado****.*

Não *se* ***fale*** *em* ***corda*** *em* ***casa*** *de* ***enforcado***

É ***falar*** *de* ***corda*** *em* ***casa*** *de* ***enforcado***

Em ***casa*** *de* ***enforcado não nomeies*** *o* ***baraço***

Em ***casa*** *de* ***enforcado não*** *se* ***fala*** *de* ***corda***

Em ***casa*** *de* ***ladrão não fales*** *em* ***baraço***

Em ***casa*** *de* ***ladrão não lembrar baraço***

all of them meaning approximately the same: 'Do not speak of rope in a hanged man's house'. Besides the fronting of the prepositional phrase (a locative), in the last for examples, notice the alternation between imperative and the use of the modal auxiliary *dever* 'should' (second sentence), the lexical variation of *corda* 'rope' and *baraço* 'string', the surprising alternation between *enforcado* 'hanged man' and *ladrão* 'thief', and the alternative use of *lembrar* 'remember' instead of *falar* 'speak'. Notice also the (truncated?) form, in the fifth line, introduced by *ser* 'to be', which can be used as a comparison and adapted for commenting on an given event/situation: *Fazer isso é* (*como*) *falar de corda* … 'to do this is like…'. In this more complex case, and because of the changes in word order, the key elements are represented by two strings:

(não+<E>) (<falar>+<lembrar>) (<corda>+<baraço>) <casa>(<enforcado>+<ladrão>)

<casa> (<enforcado>+<ladrão>) (não+<E>) (<falar>+<lembrar>) (<corda>+<baraço>)

Notice that in the case of the negation adverb *não* 'not', this should be treated as a facultative element, since the zeroing of the negation is often one of the strategies to creatively adapt the proverb to new uses. This very phenomenon is illustrated in one of the proverbs' variants.

Thus, a database in tabular format was produced, where the lines contain the proverbs keywords and the number of columns varies according to the class. This matrix, that we call a *lexicon-grammar*, can also be used to include other relevant information. For the moment, this is limited to the proverbs' conventional ID (the code of the class and the proverb number).

[45] We use the notation of the UNITEX system (Paumier 2003, 2014) in representing these regular expressions: elements inside brackets and linked by '+' are interchangeable in that given position; <E> stands for the null string.

2.3. Building finite-state tools for adverb identification

Since the lexicon-grammar of proverbs is in constant update, and can not be directly used to match strings in texts, we used UNITEX[46] linguistic development platform (Paumier 2003, 2014), which is based on finite-state technology, we built the tools that would allows to find, delimit and tag as proverbs candidate strings in running text. In this way, the linguistic information is represented independently of the tools used to apply it to texts. Next we describe this process.

A reference graph was created for each class of proverb, according to its syntactic structure and the number of keywords. Each graph describes the sequences of those key elements of the proverbs. In the graph, variables represent the corresponding cells in the matrix. Then, the reference graph is intersected with the matrix: the system reads each line of the database at a time, replacing the variables for the content of the corresponding cells in the matrix and generates a sub-graph for each proverb. Each sub-graphs is univocally numbered and entire set of is grouped together in a result graph that can be used to match proverbs in texts. To simplify the procedure, for this paper, a single reference graph was produced and a simplified extract is shown in Fig. 1:

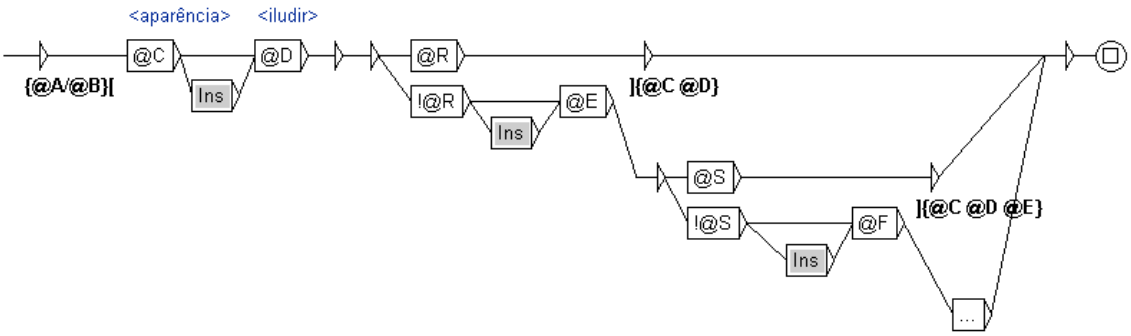

Fig. 1. Reference graph for the proverbs' database.

In this graph, variables @R and @S stand for supplementary or auxiliary columns in the matrix. These columns indicate whether the $n+1^{th}$ column has content (and then this is marked with a plus '-') or not ('+'), the n^{th} column being the last content cell that has been processed; up to 11 columns were used to represent the keywords of the proverbs; in this case, @R stands for the 3rd column, @S for the 4th, and so on; there are always at least two keywords/elements, as illustrated by the proverbial expression *As aparências iludem* 'Appearances deceive'. The system replaces the content variables in the graph (in this case, @C , @D , @E and @F) by their corresponding elements in the matrix. When reaching a variable for the auxiliary columns, the system either continues the path, if this corresponds to a plus sign '+', ending the representation of the proverb; or it moves to the next content variable. The process is repeated until all columns have been explored. Thus, a single reference graph can be used for tables with any given number of columns. The sub-graphs are transducers that can be applied to texts. When matching a given sequence, they insert a right delimiter '**[**' preceded by the proverb ID and the number of keywords (variables @A and @B); and a left delimiter '**]**' followed by the proverb's keywords (variables @C, @D, etc.). An insertion window form 0 up to 3 words is represented by an auxiliary sub-graph **Ins** (shown in a grey box). Fig. 2 shows the sub-graph automatically generated for proverb *Roma e Pavia não se fizeram num dia* 'Rome and Pavia were not made in a single day' [**P1F4**]:

[46] http://www-igm.univ-mlv.fr/~unitex/

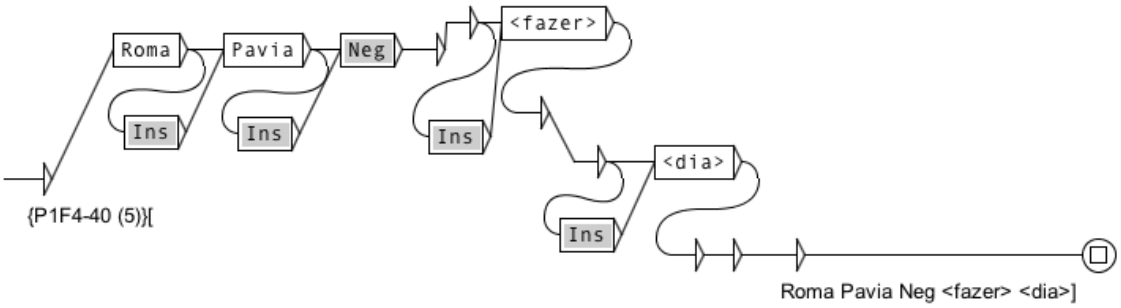

Fig. 2. Sub-graph describing an individual proverb.

2.4. Corpus building

For the automatic identification of proverbs in texts, we digitized a list of approximately 56,150 proverbs (and variants), taken from two large proverb collections:

- Costa (1999): 32,709 entries and 278,217 words;
- Machado (2011): 23,437 entries and 211,257 words.

This list will function not only as a source for the completion of the lexicon-grammar, but also as a testing ground for the finite-state tools built to identify proverbs in texts.

3. PRELIMINARY EVALUATION

The 594 sub-graphs (one for each proverb type), resulting from the process described in the previous section, were applied to the list of 56,150 proverbs (and variants) taken from the two large proverb collections mentioned above. Table 2 shows some preliminary results.

Number of proverbs	**LPP (Costa, 1999)**	**GLP (Machado, 2011)**
Proverbs in each *corpus* (matches)	# 446	# 391
Different proverbs in each *corpus* (types)	# 199	# 194
Proverbs in both *corpora* (matches)	# 347	# 310
Different proverbs in both *corpora* (types)	142	

Table 2. Number of proverbs in corpora

For the 594 proverb types represented in the graphs, only 446 and 391 were captured in each sub-corpus. Notice that this difference may be related to the fact that the list from Machado (2011) is substantially shorter than the one from Costa (1999). Notice also that in spite of the number of matches, the number of different types is much smaller, corresponding to a ratio variants/type of 2.24 in Costa (1999) and 2.02 in Machado (2011). On the other hand, the matched types correspond to about a third of all the types in the

lexicon-grammar. This can be due to several reasons: an incorrect formalization of the data in the matrix; or some problem in the graphs; or the fact that several proverbs are specific of the Brazilian variant. However, it is interesting to notice that in such large collections, so many types were of the lexicon-grammar were not found. Finally, we remark that only 142 types were common to both collections, corresponding to 347 and 310 matches, respectively. This hints at a large non-overlap of the two lists, not due to variation alone, but mostly to lexical coverage.

The most common type of variation in matched proverbs is the left and right extensions of the core elements: while the keywords of the proverb are found, the graphs also capture variants that present some extra material. This is the case, for example with proverb *Não há regra sem exceção* 'The is no rule without exception', from the class P1F1, that produced three different matches (extra material in bold):

Não há regra sem excepção

Em toda a afirmação *não há regra sem excepção*

Não há regra sem excepção, ***nem mulher sem senão***

As one can see, the first line corresponds to the basic form of the proverb, while in the following lines there is a fronted preposition phrase and coordinated clause, respectively.

Variation in vocabulary is also observed. The most affected grammatical classes are nouns and verbs, showing morphological and lexical variation but no differences regarding the proverbs meaning or function. The next two proverbs illustrate these types of variation, respectively (keywords in bold):

O ***fim justifica*** *os* ***meios***

Nem sempre os ***fins justificam*** *os* ***meios***

Os ***fins*** *não* ***justificam*** *os* ***meios***

Burro velho *não* ***aprende letra***

Burro velho *não* ***aprende línguas***

Burro velho *não* ***aprende o caminho***

Representing in the lexicon-grammar most keywords by their lemma already captures morphologic variation. However, in order to capture the lexical variation, since not all variants are yet represented in the lexicon-grammar, a good strategy would be to semi-automatically create new graphs replacing one keyword at a time by its PoS code (e.g. <N> for the nouns, <V> for the verbs, and so on), and then automatically retrieve the variants that were not yet registered in the lexicon-grammar.

4. CONCLUSION AND FUTURE WORK

This paper presented a general framework for the formal (syntactic) classification of proverbs, based on their structure and syntactic properties. The first steps towards the construction of a lexicon-grammar of Portuguese proverbs by building a database of proverb types, consisting of the keywords that univocally identify each proverb, and allowing for some lexical and morphologic variation. Preliminary experiments on the conversion of these lexical matrices into finite-state tools were carried out, which enable to identify, delimit and tag proverbs and their variants in texts. An extensive list of 56,150 proverbs (and variants), taken from two large collections was digitized and will function

not only as a source for the completion of the lexicon-grammar, but also as a testing ground for the finite-state tools built to identify proverbs in texts.

From the obvious contrast between the number of proverbs from each collection and the number of types already represented in the lexicon-grammar, the next first step is to extend the lexical coverage of the matrix. One of the methods to be explored is the semi-automatic construction of new graphs replacing each content keyword (mainly nouns, verbs and adjectives) by the corresponding PosS. As a consequence, the initial classification may need to be refined, particularly in the case of the most productive structures, which may require sub-classification or even the creation of new classes. Attention must be given to devise a reference procedure that will allow relating, in a single lexical-grammatical unit, those variants belonging to different formal classes, often involving changes in word order. A more precise understanding of the variation phenomena, and the construction of reference graphs for each formal class will reflect of a more precise definition of the insertion window and the sentence alternations that proverbs often allow. It is also a future objective, once a satisfactory lexical coverage is achieved, to measure the frequency of the proverbs types (and their variants) in real corpora, in order to associate frequency information to the database.

Acknowledgements

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

ALMEIDA, J.J. 2014. *Dicionário aberto de calão e expressões idiomáticas*. [online] Available at: <http://natura.di.uminho.pt/~jj/pln/calao/dicionario.pdf> [Accessed 14 March 2015].

BRUCKSCHEIN, M., MUNIZ, F., SOUZA, J., FUCHS, J.T., INFANTE, K., GONÇALEZ, P.N., VIEIRA, R. and ALUISIO, S.M. 2008. *Anotação linguística em XML do corpus PLN-BR. Série de Relatórios do NILC*, São Carlos (SP): NILC-ICMC-USP.

COSTA, J. R. M. 1999. *O Livro dos Provérbios Portugueses*, Lisboa: Editorial Presença.

MACHADO, J.P., 2011. *O Grande Livro dos Provérbios*, 4ª ed., Lisboa: Casa das Letras.

RASSI, A.P., 2014. *List of Proverbs in Brazilian Portuguese*. <https://www.researchgate.net/publication/269165152_Rassi2014> [Accessed 14 March 2015]. DOI: 10.13140/2.1.4907.7280

RASSI, A.P., BAPTISTA, J. AND VALE, O. 2014. Automatic Detection of Proverbs and their Variants. In: M. Pereira, J. Leal, J. and A. Simões, eds. *Proceedings of the Symposium on Languages, Applications and Technologies* (SLATE`14). Leibniz (Germany): Schloss Dagstuhl - Leibniz-Zentrum fur Informatik, Dagstuhl Publishing. pp. 235-249.

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.