Automatic Detection of Proverbs and their Variants

Amanda P. Rassi[1,2], Jorge Baptista[2], and Oto Vale[1]

1 **Federal University of São Carlos-UFSCar**
Rodovia Washington Luís, km 235 – SP-310. São Carlos – São Paulo – Brasil
CEP 13565-905
aprassi@ualg.pt,otovale@ufscar.br

2 **University of Algarve-FSCH/CECL**
Campus de Gambelas, 8005-139 Faro, Portugal
jbaptis@ualg.pt

Abstract

This article presents the task of automatic detection of proverbs in Brazilian Portuguese, from the intersection of the regular syntactic structure of proverbs and their core elements. We created finite-state automata that enabled us to look for these word combinations in running texts. The rationale behind this method consists in the fact that although proverbs may have a normal sentence structure and often a very commonly used lexicon, their specific word-combinations may enables us to identify them and their variants irrespective of the syntactic or structural changes the proverb may undergo. The goal of this task is to gather the largest number of proverbs and their variants. The results showed precision 60.15%.

1998 ACM Subject Classification I.2.7 Natural Language Processing

Keywords and phrases Brazilian Portuguese, proverbs, syntactic structure, core element, variation

Digital Object Identifier 10.4230/OASIcs.SLATE.2014.235

1 Introduction

The existence of proverbial structures in texts, including journalistic texts, is indisputable [12], which raises the problem of identifying them as a complex structure. The main problem concerning the identification of proverbs is that they have the same syntactic structure and the same words as ordinary, free sentences, however, they normally have a non compositional meaning and must be recognized not as an ordinary string of words, but as a complex unit, formed by several words, phrases and even multiple clauses. In this sense, proverbs resemble multiword expressions (MWE), although some authors [13, p.53] consider them as a different type of linguistic units as a quoted speech inside speech itself. In this paper, we adopt the view that proverbs should be treated as MWE.

In general, automatic processing of idiomatic expressions, fixed expressions, semi-fixed expressions, proverbs and other multiword expressions is still a hard task for Natural Language Processing (NLP) [30]. Although there are many studies about the identification of multiword expressions in NLP [20, 21, 23], it is still difficult to identify them automatically in natural language texts [4, 5, 26].

In this paper we focus on the special case of proverbs in view of a double problem they represent to NLP: the fact that proverbs accept both lexical and formal (structural) variation. We aim at developing a method for automatic detection of proverbs and their variants, based

3rd Symposium on Languages, Applications and Technologies (SLATE'14).
Editors: Maria João Varanda Pereira, José Paulo Leal, and Alberto Simões; pp. 235–249

OpenAccess Series in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

on existing compilations of proverbs, by exploring the regular syntactic structures that most proverbs present. These regularities led to a formal classification of proverbs, based on their syntactic structure. Finite-state automata will be used to represent the regular patterns found in these classes of proverbs. Results from the automatic identification of Brazilian Portuguese proverbs from real texts are presented. This approach can be used in to two main applications: for lexicographic work, in order to build more complete dictionaries, and for Natural Language Processing, to improve linguistic resources, tools and applications, by allowing systems to signal these micro-texts and a special type of discursive element.

2 Delimitation of the Object

Proverbs, parables, adages, aphorisms, maxims, and so on, these are all different terms used to designate similar types of sentences. Though there are conceptual differences among these terms, in practice, many authors ignore such distinctions and tend to group all these linguistic expressions under the broad umbrella term of *proverb*. In this paper, we also adopt such broad perspective and will consider proverbs as linguistic expressions forming fixed word combinations, in spite of some (limited) lexical or structural variation, often with a sentential status, that may even include subclauses, and whose global meaning is often idiomatic. These micro-texts are usually generic statements, conveying a world view or stating a moral judgement, an eternal truth, an ideal state of affairs.

We distinguish proverbs from *fixed expressions/frozen sentences* (or *idioms*, proper). In idioms, the verb and one of its argument positions are frozen together, that is, they are distributionally invariant, or the argument nouns can only vary within a small and closed paradigm. Usually the subject of frozen sentences is distributionally free, and its selection depends not just on the verb, but on the overall meaning of the combination of the verb and its frozen arguments; *i.e. Ana/Essa mesa não vale um tostão* 'Ana/This table is not worthy a penny'. On the other hand, typically, proverbs are completely frozen sentences, where, in spite of some (reduced) lexical variation and some (even more constraint) syntactical paraphrasing, all the elements are fixed. In other words, proverbs have the subject position necessarily filled by a fixed element [18, p.161], while the subject in fixed expressions usually varies and may be defined intensionally, by distributional constraints.

The second property that distinguishes proverbs and fixed expressions is, according to [24], that the proverbs "always have an autonomous semantic value in communicative terms, unlike idioms that are only constituents of sentences and may never occur as a full sentence." In this sense, proverbs take place in whole sentences while fixed expressions only replace phrases (nominal phrase, verbal phrase or prepositional phrase).

Although proverbs have syntactic structures similar to simple sentences, they can not be recognized as common sentences, but must be understood as a single block, whose syntactic slots should always be filled by specific lexical units. It means that proverbs are formed by words and phrases like any other free sentences, but they must be understood as a complex expression, a combination of words whose use is highly constraint.

When proverbs are introduced by an enunciative mark, such as *como dizem* 'as they say', *como dizia minha avó* 'as my grandmother used to say', *dizem por aí* 'people say/they say', *costuma dizer-se* 'it is often said', etc.; it is then easier to identify them because these type of marks can be extensively described. However, there is often no mark at all introducing proverbs in texts, which renders their spotting more difficult.

Finally, proverbs are prone to certain types of formal variation, particular ellipsis of one of its clause-type components, and they often undergo stylistic reformulation, in order to produce some perlocutionary effect. For example, a banking institution, in one advertisement

of its products, recently "reinvented" the proverb *Tempo é dinheiro* 'Time is money' as *Tempo não é só dinheiro. É valor* 'Time is not just money. It is value'. This capacity of the proverbs to be reinterpreted and reformulated, which some linguists called "défigement" or "unfreezing" is an inherent part of the paremiologic dynamics in language.

3 Related Works

Most of the work done on Brazilian Portuguese proverbs adopt a didatic or pedagogic approach, [14, 25, 31], or analyzes rhetorical relations between the clauses [15, 16, 17]. We did not find any work that describes formally proverb structures in Portuguese or that tried to identify them automatically in large corpus.

For European Portuguese, Lucília Chacoto developed many studies on proverbs, either theoretical and practical works. The author compared Portuguese and Spanish proverbs initiated by *Quem/Quien* 'Who' [6] and also analyzed comparative structures [7] which are two of the structures we describe in this paper.

We can also cite works for other languages, like Lacavalla [22], who compared proverbs initiating by *Quand/Quando* 'When' in Italian and French. The author uses local grammars for searching the proverbs in both languages and describes the data in Lexicon-Grammar Tables, analyzing all syntactic properties and distribution of those units. On the other hand, Navarro Brotons [2] compared proverbs in Spanish and French. The author analyzed syntax, semantics and translation of proverbs and their variants in both languages and also described the data in Lexicon-Grammar tables.

We also cite the extensive work of Mirella Conenna [8, 9, 10, 11], who produced many works about proverbs in French and Italian, comparing their structures in both languages, classifying proverbs in syntactic tables, *i.e.* Lexicon-Grammar tables, and analyzing proverbs and their variants in equivalence classes. In all those works, the author was concerned about the formalization of the data for automatic identification and processing.

There are also some other publications about proverbs in Brazilian Portuguese, but they do not present any systematic analysis. These include didactic materials used in schools, dictionaries, glossaries, and lists of proverbs. Most of them are used in teaching/learning Portuguese as second language or as didactic manuals.

For Brazilian Portuguese it is still necessary to describe formally syntactic structures of the proverbs and their core elements, aiming to contributing for the construction of lexicon-syntactic resources applicable in NLP.

4 Methods

In this section we present a methodology for automatic detection of proverbs and their variants, tested on a Brazilian Portuguese corpus, which can be resumed in 6 steps: (i) creating a database with proverbs searched in dictionaries and other lists; (ii) defining syntactic criteria to organize the collected proverbs into formal classes; (iii) manually identifying the POS tags of their elements; (iv) generating tables with the core elements derived from POS tagging; (v) creating graphs with the basic structure for each class; and (vi) intersecting the graphs with the tables of the proverbs' core elements to produce finite-state transducers that will enable us to identify such word combination in texts. After these steps, we could find other proverbs and their semantic variations within the same syntactic structure.

We searched for the proverbs and their variants in PLN.BR Full corpus [3], which contains 103,080 texts, with 29,014,089 tokens, from *Folha de São Paulo*, a Brazilian newspaper, from 1994 to 2005.

4.1 Collection of Proverbs

The first step for this work consists in creating a list of proverbs that will serve as input seeds to recognize other proverbs and their variants in large corpora. Five different sources were used: a list of proverbs in Wikipedia, three books with proverbs collections [29, 32, 34] and a dictionary of proverbs [19].

Firstly, all the expressions collected in these sources were analyzed manually and many were discarded as they were not considered as proverbs but consist mostly of idiomatic expressions (or idioms), like (1), or aphorisms and maxims, as in (2):

(1) *Matar dois coelhos com uma cajadada só*
[to] kill two bunnies with just one thwack 'kill two birds with a stone'

(2) *Na natureza, nada se cria nada se perde, tudo se transforma*
'In Nature, nothing is created, nothing is lost, everything is transformed'

The idiom in (1) is a frozen sentence with a free subject slot and two frozen complements, a direct object and an instrumental complement [1, 18, 35](class C1P2). On the other hand, (2) is an aphorism or maxim, attributed to the chemist Lavoisier (1743-1794) about the conservation of mass. In spite of its three-clause, parallelistic, proverb-like structure, and its generic nature, the (known) authorship of the maxim lead us to discard it from our study.

After a substantial collection of over 3,502 proverbs (and their variants) has been gathered, the variants of each proverb were grouped together and one of them was selected to be considered as the entry of our lexicon (or its base-form), based on its frequency among the sources consulted. Most differences between variants of the same proverb consist in the variation of their grammatical elements, and the lexical choices for their core meaningful words.

Finally, we tried to confirm whether these proverbs were (still) really in use in current Brazilian Portuguese, checking them with 5 native speakers of Brazilian Portuguese from different geographic regions.[1] Some proverbs are only used in Portugal or in Portuguese-speaking African countries, while others are very old and probably may not be in use anymore.

From the original 3,502 proverbs (and their variants), a final list of 594 proverbs (*types* or *base-forms*) was compiled.[2]

4.2 Classifying Proverbs and POS Tagging their Elements

The list of proverbs (base-forms) was then classified into formal classes. This classification was based on the following criteria, applied in this order:

(i) the number of propositions (one, two, or three clauses or clause-like units);
(ii) coordination (in multiple-clause proverbs);
(iii) order of the main vs the subordinate clauses (in multiple-clause proverbs);
(iv) order of the constituents (in single-clause proverbs);
(v) impersonal constructions; and
(vi) obligatory negation.

Table 1 presents the current classification.

[1] We consider that the sampling by region is not sufficient to confirm the presence or absence of proverbs, and we would need to consult speakers from different genders, ages, social classes, education levels etc, this is out of the main scope of this work.

[2] The list of proverbs and their classification can be consulted at the first author profile in ResearchGate, available in `https://www.researchgate.net/project/PB-proverbs`.

Table 1 Formal Classification of Brazilian Portuguese Proverbs.

Class	Structure	Example (approximate translation)	Types
P1F1	*Ø V w* (impersonal)	*Não há crime sem lei* 'There is no crime without law'	20
P1F2	N_0 *Vcop Adj/N w*	*A carne é fraca* 'The flesh is weak'	53
P1F3	N_0 *V w*	*O hábito (não) faz o monge* 'The cloth (does not) make the monk'	80
P1F4	N_0 *Neg V w*	*Burro velho não aprende línguas* 'Old donkey does not learn languages'	53
P1F5	*Prep* N_i N_0 *V w* (fronted prep. phrase)	*Para bom entendedor, meia palavra basta* 'For the one who understands, half word is enough'	45
P2F1	F_1 *Conjs-comp* F_2 (comparatives)	*Mais vale um pássaro na mão do que dois voando* 'Beter is is a bird in the hand than two flying'	39
P2F2	F_1 *Conjc* F_2 (coordinated)	*A palavra é de prata e o silêncio é de ouro* 'The word is silver and the silence is gold'	71
P2F3	N_1, N_2	*Tal pai, tal filho* 'Like father, like son'	48
P2F4	*Qu-* F_1 F_2 (interrogative subclass)	*Quem tem boca vai a Roma* 'Who has a mouth goes to Rome'	90
P2F5	F_1 *Conjs* F_2 (subordinated)	*Os amigos são muitos quando grande é a abastança* 'Friends are many when abundance is great'	20
P2F6	*Conjs* F_2, F_1 (fronted subord.)	*Quando a esmola é demais, o santo desconfia* 'When alms are too much, the saint gets suspicious'	28
P3	F_1, F_2, F_3	*Um é pouco, dois é bom, três é demais* 'One is little, two is good, three is too much'	47
Total			**594**

Some remarks on this classification are in order:

(i) impersonal constructions involve the verb *haver* 'there be' and *ter* 'to have' with impersonal valency (the later only exists in Brazilian Portuguese);

(ii) sentences with copula verbs *ser* and *estar* 'to be' usually present an adjectival or nominal predicate; these sometimes allow for mirror permutation (*A carne é fraca* = *fraca é carne*[3] 'The flesh is weak');

(iii) proverbs with obligatory negation usually involve negation adverbs, *e.g. não* 'no/not', *nunca* 'never', *jamais* 'never', *nem* 'nor', etc.; negation has precedence over copula verbs, so that proverbs with negated copula were included in this class;

(iv) single-clause proverbs with a fronted prepositional phrase do not admit the basic word-order;

(v) comparative proverbs, including those with subordinate sub-clause, are a type of complex sentences, though other types of comparative structures were also included in this class;

(vi) nominal propositions named N_1, N_2 (in P2F3 class) are treated as clausal propositions, even if they may contain no verbs and only have a 'clausal' or 'propositional content'.

[3] `http://rainhadocarmelo.blogspot.pt/2010_02_01_archive.html` [2014-03-08 13:11]

After classifying the proverbs, we manually annotated their elements for part-of-speech (POS) tags. Since each class is syntactically homogeneous, it was then relatively simple to organize the lexical items in a tabular format, so that the characteristic elements of the proverbs may be aligned, and can easily be identified. For the noun phrases (*NP*), either the subject (N_0) or the complement (N_1), the head noun (or pronoun) is determined, and eventual determiners (*Det*) or modifiers (*Mod*) are tagged and distributed across the corresponding columns. Eventual pre- or post-modifiers of verbs (*Deus escreve* direito *por linhas tortas* 'God writes straight with crooked lines'), including obligatory auxiliary verbs (*Não se entra em briga que não se* pode *ganhar* 'Do not enter into a fight you can not win'), and other elements, such as the impersonal pronouns (*Aqui se faz, aqui se paga* 'Here you do, here you pay')[4], or obligatory negation (*Quem não tem cão caça com gato* 'Who does not have a dog hunts with a cat') are also taken into consideration. Subordinative or coordinative elements are also provided with an adequate slot. In this way, it is relatively simple to automatically extract the core (or more representative) elements from each proverb, based on the classes' formal homogeneity.

4.3 Extracting Core Elements

In order to extract the core words in each proverb, we analyzed all cells in each table and selected as core elements the most frequent grammatical classes in each syntactic position. For example, in almost all classes[5] the initial *NP* is necessarily filled by a noun or, in rare cases, a pronoun. The noun can be accompany by determinants and/or adjectives and/or other nominal adjuncts, but the only position that is fully filled by some element is the column <N> either in the subject or in the complement position, so we selected the item instantiated in column <N> as one of the core elements for identifying the proverb.

In all classes[6], *VP* position is necessarily filled by a verb, so this is selected as a key element in the constitution of the proverbs. Table 2 shows a sample of P1F3 class, in a tabular format, indicating all columns[7].

Depending on the formal class of the proverbs, so the core elements are defined. In the case of class P1F2, the definitory elements are the heads of the subject and of the predicative complement (noun or adjective) as well as the copula verb. In the case the head in null (*e.g. Os últimos serão os primeiros* 'The first shall be the last') the determiner or an adjective may be chosen instead. In comparative proverbs, there is often no main verb, so the determiners 4.3 or the comparative conjunctions 4.3 must be selected, along with the core nouns:

(3) ***Tal*** *pai* ***tal*** *filho*
'**Like** father **like** son'

(4) ***Nem tanto*** *ao mar* ***nem tanto*** *à terra*
'**Not so much** to sea **not so much** to ground'

---

[4] In Portuguese, impersonal clitic pronoun *-se* imposes 3^{rd} person-singular agreement to the verb, thus being indistinguishable from passive-like pronominal constructions. Only some few clear-cut cases of pronominal passives were found; e.g. *Entre mortos e feridos salvaram-se todos* 'Among dead and wounded all were saved'. Both strategies may be considered as a form of subject (agent) degenerescence, hence contributing to the generic effect of the proverbs.

[5] Exception done for class P1F1, which has no explicit subject (null subject).

[6] Exception done for class P2F3, which is constituted by nominal phrases only, and has no verb.

[7] In this table the headings are read as follows: Adj = Adjective, Adv = Adverb, Det = Determinant, Indet_Pass = Pronominal passive-like construction, N = Noun, Prep = Preposition, V = Verb; the words inside chevrons correspond to lemmas

Table 2 Sample of class P1F3.

Proverb		Det	Adj	N	Adj	Indet_Pass	V	Adv	Prep	Det	Adj	N	Adj
A adversidade faz os heróis	<o>	-		<adversidade>	-	-	<fazer>	-	-	<o>	-	<herói>	-
A ambição cega a razão	<o>	-		<ambição>	-	-	<cegar>	-	-	<o>	-	<razão>	-
A intenção faz o agravo	<o>	-		<intenção>	-	-	<fazer>	-	-	<o>	-	<agravo>	-
A justiça começa em casa	<o>	-		<justiça>	-	-	<começar>	-	em	-	-	<casa>	-
A ocasião faz o ladrão	<o>	-		<ocasião>	-	-	<fazer>	-	-	<o>	-	<ladrão>	-
A união faz a força	<o>	-		<união>	-	-	<fazer>	-	-	<o>	-	<força>	-
As aparências enganam	<o>	-		<aparência>	-	-	<enganar>	-	-	-	-	-	-
As más notícias chegam depressa	<o>	<mau>		<notícia>	-	-	<chegar>	depressa	-	-	-	-	-
As paredes têm ouvidos	<o>	-		<parede>	-	-	<ter>	-	-	-	-	<ouvido>	-
Boas contas fazem bons amigos	-	<bom>		<conta>	-	-	<fazer>	-	-	-	<bom>	<amigo>	-
Deus escreve certo por linhas tortas	-	-		<deus>	-	-	<escrever>	certo	por	-	-	<linha>	<torto>
Mentira tem perna curta	-	-		<mentira>	-	-	<ter>	-	-	-	-	<perna>	<curto>
Muitos cozinheiros estragam a sopa	<muito>	-		<cozinheiro>	-	-	<estragar>	-	-	<o>	-	<sopa>	-
O abismo atrai o abismo	<o>	-		<abismo>	-	-	<atrair>	-	-	<o>	-	<abismo>	-
O hábito faz o monge	<o>	-		<hábito>	-	-	<fazer>	-	-	<o>	-	<monge>	-
O justo paga pelo pecador	<o>	-		<justo>	-	-	<pagar>	-	por	<o>	-	<pecador>	-
O peixe se conhece pela boca	<o>	-		<peixe>	-	se	<conhecer>	-	por	<o>	-	<boca>	-
Os fins justificam os meios	<o>	-		<fim>	-	-	<justificar>	-	-	<o>	-	<meio>	-
Roupa suja se lava em casa	-	-		<roupa>	<sujo>	se	<lavar>	-	em	-	-	<casa>	-

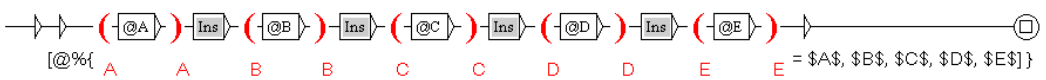

Figure 1 Reference graph for class P2F4.

In the common cases where a lexical element of the proverb allows for variation, all the variants are included in the corresponding slot. This is the case of the proverb *Cachorro mordido de cobra tem medo de linguiça* 'Dog bitten by a snake is afraid of sausage' where the second noun can be replaced by *barbante* 'string' and *salsicha* 'sausage'; notice, however, that the variation of grammatical elements 4.3 was ignored:[8]

(5) *Cachorro (que foi + <E>) mordido (de + por) cobra tem medo até de (barbante + salsicha + linguiça)*
'Dog (that was + <E>) bitten by a snake is afraid of (string + sausage + pork sausage)'

4.4 Creating and Applying the Graphs

Once the characteristic elements of each proverb have been identified, they were structured in a tabular format, one table for each class (residual class "others" was not considered in this paper). Then, using the Unitex 3.1.beta linguistic development platform [27, 28], we produce a reference graph for each class. Fig. 1 illustrates the graph for class P2F4, corresponding to proverbs with a fronted subordinated clause; e.g. *Se queres conhecer o vilão, põe-lhe um pau na mão* 'If you want to know a villain, put a stick in his hand'.

This graph reads as follows: the system explores systematically each line in the table of a class core elements, replacing the variables *@A*, *@B*, etc, by the corresponding content of columns A, B, etc. These input variables are then associated to output variables (in the letters below the brackets) to be reused in the output. In this case, the graph delimits the matched expression by brackets, and produced the content in a normalized form, introduced by the idiom number (the table's line number), represented by variable @%[9]. By intersecting the reference graph with the corresponding table, the system generates one subgraph for each line of the table, and a general result graph, containing all the subgraphs. The result graph can then be used to find patterns in texts. Table 3 shows a sample of a concordance of such matched strings from the PLN.Br corpus.

Each line in the table has been numbered. In this concordance, a small left context is provided, followed by the number of the proverb type in the corresponding class, the actual words in the corpus and the core words that the transducer detected; empty variables are not represented (void commas).

The table presents two matches that are considered False Positives, in lines 16 and 17. The proverb supposed to be found is *Quem sabe faz* 'Who knows makes', but the system found, for example, a free sentence (line 16) and a verse of a brazilian song (line 17). It is also remarkable the transformations (actualizations or adaptations) created by speakers. The proverb we were looking for is *Quem vê cara não vê coração* 'Who sees the face does not see the heart' as in line 22, but the speaker adapted the proverb to the context of smoking and created *Quem vê cara não vê pulmão* 'Who sees the face does not see the lung', as

[8] The items linked by "+" inside parentheses can comute in the given syntactic slot; the symbol <E> represents the empty string.
[9] The shadowed box **Ins** is a subgraph defining a window of 0 to 3 words and separators allowed between the proverbs' core elements.

Table 3 Sample of a concordance of Class P2F4.

1	é o	[0003	barato que pode sair caro=barato, caro„,]
2	não	[0006	mata engorda=mata, engorda„,]
3	Quem	[0015	avisa amigo é=avisa, amigo„,]
4	Quem	[0018	cala consente=cala, consente„,]
5	Quem	[0019	Canta Seus Males Espanta=Canta, Males, Espanta„]
6	e como	[0020	casei e quero casa=casei, quero, casa„]
7	quem	[0023	conta um conto aumenta um ponto=conta, conto, aumenta, ponto,]
8	quem	[0028	diz o que quer ouve o que não quer=diz, quer, ouve, quer,]
9	não	[0042	arrisca não só não petisca=arrisca, petisca„,]
10	que não	[0043	choram nem mamam=choram, mamam„,]
11	não	[0044	deve não teme=deve, teme„,]
12	Quem	[0047	está dentro quer sair e quem está fora não=está, dentro, quer, sair,]
13	não	[0050	sabe não ensina=sabe, ensina„,]
14	quem	[0062	pariu Mateus que o embale=pariu, Mateus, embale„]
15	quem	[0064	procura acha=procura, acha„,]
16	Quem	[0068	sabe alguém faz uma experiência com isso=sabe, faz„,]
17	quem	[0068	sabe faz a hora=sabe, faz„,]
18	Quem	[0068	Sabe Faz ao Vivo=Sabe, Faz„,]
19	Quem	[0069	sabe sabe=sabe, sabe„,]
20	os que	[0070	semeiam ventos colhem tempestades=semeiam, ventos, colhem, tempestades,]
21	"Quem	[0079	tem pressa come cru=tem, pressa, come, cru,]
22	"quem	[0085	vê cara não vê coração=vê, cara, coração„]
23	quem	[0085	vê cara não vê pulmão=vê, cara, vê„]
24	Quem	[0085	vê cara vê muito mais do que coração=vê, cara, vê, coração,]
25	Quem	[0086	viver verá=viver, verá„,]

in line 23. In 24 the obligatory negation of the original proverb has been deleted and the meaning actually inverted in a creative way.

In this way it was possible to find other variants of proverbs than those we had previously collected (from books, dictionaries and the wikipedia) and find several instances of creative reuse and transformations of proverbs for rethoric purposes.

5 Results and Discussion

Since, to our knowledge, there is no available corpus annotated with proverbs and similar expressions, only precision was reported here.

From the previous list of 594 proverbs, 788 matches were found in the PLN.Br corpus, from which 474 matches (60.15%) correspond to actual proverbs. We decided to search these lexical units in journalistic corpus aiming to check if in the common language they also appear. It has been proved [33] that literary corpora contain a large number of proverbs, but the challenge is looking for them in non-literary texts. Table 4 shows the breakdown of these results by class. In spite of the number of matches, only 137 types (different proverbs) were found. The scarcity of the occurrence of proverbs in the corpus (1:36,820 words), as well as its reduced variety (23% types) is most probably linked to the journalist nature of the corpus.

In this respect, it is remarkable the number of instances retrieved from the data in class P2F4 as well as its low precision (27.5%). This class includes only two lexical items, besides the indefinite subject pronoun *quem* 'who', as in *Quem cala consente* '[he] who silence [gives

Table 4 Results of automatic identification of proverbs by class.

Class	Proverbs (types)	Matches	Types	False-Positives
P1F1	20	15	4	2
P1F2	53	91	21	16
P1F3	80	153	24	55
P1F4	53	61	15	0
P1F5	45	63	5	6
P2F1	39	40	7	1
P2F2	71	14	3	9
P2F3	48	40	8	25
P2F4	90	276	37	200
P2F5	20	3	1	0
P2F6	28	1	1	0
P3	47	31	11	0
Total	**594**	**788**	**137**	**314**

his] consent'. Since these are very short proverbs, a window of 5 words between the core elements may be inadequate.

We repeated the experiment without any insertion window, and captured 56 matches, of which 26 were false positives. The local precision of the class P2F4 raised from 27.5% to 53.57%. Considering the global precision (including all classes), global precision raised from 60.15% to 73.35%. This may indicate that, depending of the syntactic structure of the proverb, a more or less wide window between the core elements must be defined.

The system matched 137 different proverbs from the previous list with 594 entries, and their distribution is presented in Fig. 2, below. Some few other proverbs have higher frequencies but they were collapsed in Fig. 2 because they form a small number of proverbs with relatively high frequency.[10]

The small number of different proverbs matched by the system (23% of the total types) is probably due to the nature of the corpus. Some proverbs, as we will see below, have been adapted and reconfigured to fit the discursive needs of the author.

[10] Namely, f=13, f=16, f=20, f=22, f=44, f=52, f=55 and f=88.

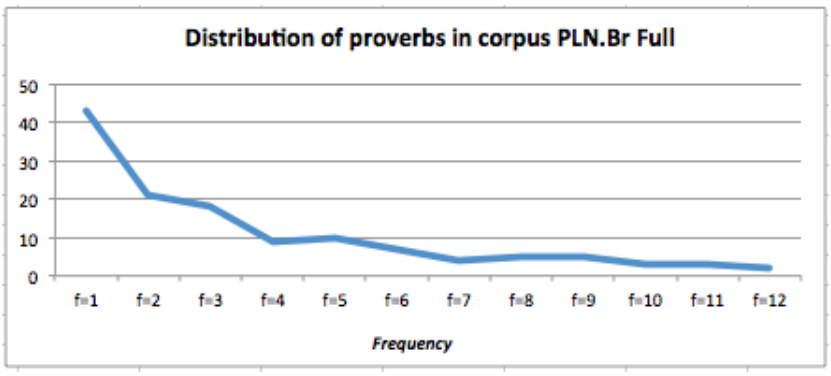

Figure 2 Distribution of proverbs in corpus PLN.Br Full.

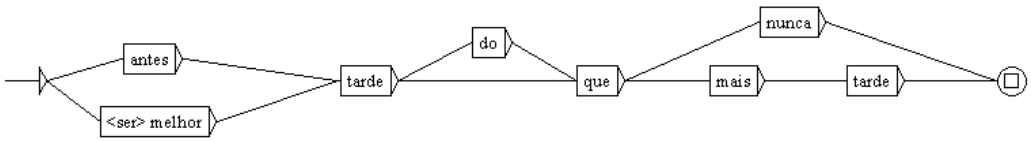

Figure 3 Graph with variants of the proverb *Antes tarde do que nunca* 'Better later than never'.

The matches found allowed us to identify other variants of the same proverb that were not in the initial list. For example, along the form *Antes tarde do que nunca* 'Better later than never', the variants can be represented by the graph presented in Figure 3.

It was also possible to find proverbs that were not in the previous list. For example, we used the structure [*quem V V*] ['who V V'], which was searched in Unitex by the following regular expression: *quem* (<MOT>+<E>)(<V:P3s>+<V:J3s>) (<MOT>+<E>) <V:P3s>. This syntaxe means: pronoun *quem* followed by a verb in the third singular person of the verb in simple present or simple past, which is followed by a verb in simple present in third singular person; between these elements a single, facultative word could also appear. This regular expression could be instantiated by *Quem sabe faz* 'Who knows makes' 5 and another similar syntactic structure was found 5:

(6) *Quem sabe faz*
'Who knows makes'

(7) *Quem sabe faz ao vivo*
'Who knows makes it viva'

These are two different proverbs, not only variants, because their meanings are different, so the task is also valid for searching more proverbs.

While the definition of the core elements is basically a lexical decision, the length of the insertion window between them is a matter of empirical decision, and it can vary, as we have seen, depending on the type of proverb involved. Several tests were conducted with insertion windows of different lengths, and, in general, results fell rapidly when more than 5 words could be inserted. The two examples 5–5, below, show 5 words between the core elements.

(8) *o buraco* [das negociações com o Congresso] *é muito mais embaixo*
'the hole [in negotiations with Congress] is much more down'

(9) *a justiça* [que o brasileiro tanto almeja] *começa dentro de casa*
'the justice [that the Brazilian so much craves] begins at home'

Another issue that had to be considered in the insertion window is the fluctuation of punctuation marks. In Portuguese proverbs, the use of comma is not systematic, and in many cases it can be considered to be optional. Particularly, in verse-like proverbs, with parallel metric in each hemistich, an hyphen '–' or even a slash '/' can be found. The reference graphs allow the facultative presence of punctuation between the core words of the proverb so that both forms are retrieved; *e.g.* 5–5:

(10) *Quem sai ao vento (,) perde o assento* (comma facultative)
'Who leaves to the the wind, loses the seat'

(11) *Quando a esmola é demais (,) o santo desconfia* (comma facultative)
'When the alms are too much, the saint suspects'

The lemmatization of the core words also raises several interesting issues. Many words were lemmatized aiming to identify all inflected forms of the verbs and the nouns, but for proverbs with the structure [*V Cop V*], such as *Recordar é viver* 'To remember is to live',

Amar é sofrer 'To love is to suffer', *Querer é poder* 'To want is to be able', among others, only the infinitive can be used, so we decided that the surface form should appear in the lexicon-grammar table.

Some proverbs admit transformations. For example, almost every proverb in class P1F2 allows the mirror permutation, which consists in reversing the order of constituents (subject and predicative) around the copula verb *ser* 'to be'; *e.g.* 5–5:

(12) *O ataque é a melhor defesa* [Mirror Perm.]= *A melhor defesa é o ataque*
'The attack is the best defense = The best defense is the attack'

(13) *A fome é o melhor tempero* [Mirror Perm.]= *O melhor tempero é a fome*
'Hunger is the best seasoning = The best seasoning is hunger'

The mirror permutation was only found in proverbs with a *NP* in the predicative position. In the case of adjectival structures, as in the proverbs *A carne é fraca* 'The flesh is weak', *O amor é cego* 'Love is blind' and *Errar é humano* 'To make mistake is human', this transformation is more rarely observed, though it can still be found in the web, so we extended it to the entire set of this class:

"*Quão fraca é a carne humana!*" [11];
"*O que você quis dizer com "Eu não sabia o quão cego é o amor."?*" [12];
"*Eu a amo, já relevei mtas coisas, mas humano é errar, burrice é repetir os erros. Cansei.*"[13]

Class P1F4 was distinguished from P1F2 and P1F3 because of the presence of an obligatory negation element, such as *não* 'not', *nunca* 'never', *jamais* 'never', among others. However, wordplay often involves the removal of this negation, to produce some type of effect. For example, on par with the proverb *Beleza não põe mesa* 'Beauty does not set the table', an affirmative variant 5 was found in the corpus :

(14) *Como a maioria das outras entrevistadas, Astrid diz que beleza põe mesa, sim*
'Like most other interviewees, Astrid says that beauty does set the table, yes'

Naturally, the interpretation of this sentence implies the previous knowledge of the negative form of the proverb. However, because of this creative re-use of the negative structure, the negation element was not considered an obligatory core element of the proverb.

Class P2F2 consists of 71 proverbs, formed by two coordinated propositions. Many of them result from the sum of two simple proverbs with one proposition each, *e.g.* the proverb 5 results from the combination of the proverbs 5 and 5, so it is considered a proverb and not just a variant.

(15) *Quem casa não pensa, quem pensa não casa*
'Who gets married doesn't think, who think doesn't get married'

(16) *Quem casa não pensa*
'Who gets married doesn't think'

(17) *Quem pensa não casa*
'Who think doesn't get married'

In these cases, in which a proverb is formed by two clauses, but also admits that only one of the clauses be used independently, the proverb was inserted thrice: in P2F1 class or in P2F2 (two clauses), and in P1F3 or P1F4 classes (single clause classes).

[11] `http://www.pastoralis.com.br/pastoralis/html/modules/newbb/` [2014/03/23]
[12] `http://m.fanfiction.com.br/reviews/historia/58620/capitulo/439083` [2014/03/23]
[13] `http://www.segredototal.com.br/de/homem/` [2014/03/23]

6 Final Remarks

In this paper we presented a methodology for detecting proverbs automatically in running texts. Proverbs have a similar syntactic structure and contain the same lexicon as ordinary free sentences, but they must be interpreted as a single unit of meaning. However, they often lack the presence of introductory expressions, that signal them as quotations, or are recast (and reshaped) in the ordinary stream of discourse, so it is necessary to recognize them in texts as multiword meaning units at a sentential/clausal level.

The results of this study showed contributions both for theoretical linguistics and to automatic text processing. As linguistic contributions, we emphasize:

(i) the formal (syntactic) classification of proverbs in 12 classes; this classification may serve as a starting point for deeper analysis on each one of these proverbial structures, as it has been done for the Spanish, French and Italian [2, 10, 11, 22];

(ii) the identification of the core elements of each proverb; the methodology presented to extract keywords can be replicated for other different *corpora* in order to see if the results are consistent across the different text types and domains;

(iii) the definition of an adequate extent of a window for insertions (words and punctuation), which may vary depending on the formal class; and

(iv) the frequent occurrence of variation, including of transformational nature, such as the mirror-permutation, and the zeroing of negation elements.

As contributions for automatic processing of texts in natural language, we highlight:

(i) the evaluation of the task, which showed 60.15% of precision with a 0-5 words window and 73.35% when no insertion is allowed; and

(ii) the construction and application of reference graphs for automatic detection of the proverbs and their variants in large corpus.

Naturally, much is still to be done.

Acknowledgements. This work was partially supported by national funds through FCT – Fundação para a Ciência e a Tecnologia, under project PEst-OE/EEI/LA0021/2013 and by Capes/PDSE under Process BEX 12751/13-8. We would like to thank the comments of the anonymous reviewers, which helped to improve this paper.

References

1 Jorge Baptista, Anabela Correia, and Graça Fernandes. Léxico-gramática das frases fixas do portugués europeo. *Cadernos de Fraseoloxía Galega*, pages 41–53, 2005.

2 María Lucía Navarro Brotons. *Las paremias y sus variantes: análisis sintáctico, semántico y traductológico español/francés*. PhD thesis, Universidad de Alicante, Alicante, Spain, 2008.

3 M. Bruckschein, F. Muniz, J. G. C. Souza, J. T. Fuchs, K. Infante, M. Muniz, P. N. Gonçalez, R. Vieira, and S. M. Aluisio. Anotação linguística em xml do corpus pln-br. Série de relatórios do nilc, NILC – ICMC – USP, 2008.

4 Lars Bungum, Björn Gambäck, André Lynum, and Erwin Marsi. Improving word translation disambiguation by capturing multiword expressions with dictionaries. In *Proceedings of the 9th Workshop on Multiword Expression*, pages 21–30, Atlanta, Georgia, USA, June 2013.

5 Helena M. Caseli, Carlos Ramisch, Maria das Graças Volpe Nunes, and Aline Villavicencio. Alignment-based extraction of multiword expressions. *Language Resources and Evaluation – Special Issue on Multiword expression: hard going or plain sailing.*, pages 59–77, 2010.

6 Lucília Chacoto. A sintaxe dos provérbios – as estruturas quem/quien en portugués e español. *Cadernos de Fraseoloxía Galega*, pages 31–53, 2007.

7 Lucília Chacoto. Mais vale mais um gosto na vida que três vinténs na algibeira – las estructuras comparativas en los proverbios portugueses. *Aspectos formales y discursivos de las expresiones fijas*, pages 87–103, 2008.

8 Mirella Conenna. Acerca del tratamiento informático de los proverbios. *Léxico y fraseología*, pages 197–204, 1998.

9 Mirella Conenna. Sur un lexique-grammaire comparé de proverbes – les expressions figées. *Langages*, 90:99–116, 1998.

10 Mirella Conenna. Classement et traitement automatique des proverbes français et italiens. *Lexique, Syntaxe et Sémantique, Mélanges offerts à Gaston Gross à l'occasion de son soixantième anniversaire*, pages 285–294, 2000.

11 Mirella Conenna. Dictionnaire électronique de proverbes français et italiens. In *Actes du XXIIe Congrès International de Linguistique et de Philologie Romanes*, pages 137–145, Bruxelles, Juillet 2000.

12 Mirella Conenna. Principes d'analyse automatique des proverbes. *Syntax, Lexis & Lexicon-Grammar, Papers in honour of Maurice Gross*, pages 91–103, 2004.

13 Paul Cook and Graeme Hirst. Automatically assessing whether a text is cliched, with applications to literary analysis. In Valia Kordoni, Carlos Ramisch, and Aline Villavicencio, editors, *Proceedings of the 9th Workshop on Multiword Expression*, pages 52–57, Atlanta, Georgia, USA, June 2013. Association for Computational Linguistics.

14 Márcia de Carvalho Saliba. Unidades lexicais maiores que a palavra: descrição linguística, considerações psicolinguísticas e implicações pedagógicas. Master's thesis, Universidade Federal do Paraná, Paraná, 2000.

15 Ana Clara Gonçalves Alves de Meira. Uma análise da articulação de cláusulas hipotáticas adverbiais em provérbios do portugês brasileiro. In EDUFU, editor, *Anais do SILEL*, volume 1, Uberlândia-UFMG, 2009.

16 Ana Clara Gonçalves Alves de Meira. A articulação de orações em provérbios do portugês em uso: uma análise das relações retóricas. Master's thesis, Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, 2011.

17 Glaucy Ramos Figueiredo. *O gênero proverbial na imprensa: usos e funções retóricas*. PhD thesis, Universidade Federal de Pernambuco, Recife-PE, 2012.

18 Maurice Gross. Une classification des phrases figées du français. *Révue Québécoise de Linguistique*, 11(2):151–185, 1982.

19 Raimundo Magalhães Jr. *Dicionário brasileiro de provérbios, locuções e ditos curiosos: bem como de curiosidades verbais, frases feitas, ditos históricos e citações literarias, de curso corrente na língua falada e escrita*. Documentário, Rio de Janeiro, 3 ed edition, 1974.

20 Valia Kordoni, Carlos Ramisch, and Aline Villavicencio, editors. *Proceedings of the ACL Workshop on Multiword Expressions: from Parsing and Generation to the Real World (MWE 2011)*, Portland, OR, USA, June 2011.

21 Valia Kordoni, Carlos Ramisch, and Aline Villavicencio, editors. *Proceedings of the 9th Workshop on Multiword Expression*, Atlanta, Georgia, USA, June 2013.

22 Cláudia B. Lacavalla. *Lexique-grammaire des proverbes en Quand/Quando – Comparaison français-italien et représentation par grammaires locales*. PhD thesis, Universitá degli Studi di Bari, Bari, Itália, 2007.

23 Éric Laporte, Preslav Nakov, Carlos Ramisch, and Aline Villavicencio, editors. *Proceedings of the COLING Workshop on Multiword Expressions: from Theory to Applications (MWE 2010)*, Beijing, China, August 2010.

24 Ana Cristina Macário Lopes. *Texto Proverbial Português – Elementos para uma análise semântica e pragmática*. PhD thesis, Universidade de Coimbra, Coimbra, 1992.

25 Maria Lucia Mexias-Simon. Para uma estrutura dos provérbios nas línguas românicas: uma experiência. *Mosaico – Revista Multidisciplinar de Humanidades*, 2(2):59–74, 2011.

26 Martha Palmer. Complex predicates are multi-word expressions. In *Proceedings of the 9th Workshop on Multiword Expression*, page 31, Atlanta, Georgia, USA, June 2013.

27 Sébastien Paumier. *De la reconnaissance des formes linguistiques à l'analyse syntaxique.* PhD thesis, Université de Marne-la-Vallée, 2003.

28 Sébastien Paumier. *Unitex 3.1 – Manuel d'Utilisation*, last edition, 2013.

29 Ciça Alves Pinto. *Livro dos provérbios, ditados, ditos populares e anexins.* Senac, São Paulo, 4 ed edition, 2003.

30 Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. Multiword expressions: A pain in the neck for NLP. In *Proc. of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002)*, pages 1–15, 2001.

31 Ana Paula Gonçalves Santos. Análise da escolha lexical no estudo dos provérbios em LP. In *Anais do SIELP*, Uberlândia-UFMG, 2012. EDUFU.

32 Martha Steinberg. *1001 provérbios em contraste: provérbios ingleses e brasileiros.* Editora Ática, São Paulo, 1985.

33 José Teixeira. Mecanismos metafóricos e mecanismos cognitivos: Provérbios e publicidade. In Arco Libros, editor, *Actas del VI Congreso de Lingüistica General*, pages 2271–2280, Madri, 2007.

34 Nelson Carlos Teixeira. *O grande livro de provérbios.* Leitura, Belo Horizonte, 1942.

35 Oto Araújo Vale. *Expressões cristalizadas do português do Brasil: uma proposta de tipologia.* PhD thesis, Universidade Estadual Julio Mesquita Filho – UNESP, 2001.