diacrítica
REVISTA DO CENTRO DE ESTUDOS HUMANÍSTICOS
Vol. 33, n.º 2, 2019, p. 324-334. DOI: doi.org/10.21814/diacritica.412

DESCRIÇÃO E CLASSIFICAÇÃO SINTÁTICA DAS EXPRESSÕES PROVERBIAIS DO PORTUGUÊS EUROPEU

DESCRIPTION AND SYNTACTIC CLASSIFICATION OF EUROPEAN PORTUGUESE PROVERBS

Sónia Reis[*]
reis.soniamm@gmail.com

Jorge Baptista[**]
jbaptis@ualg.pt

Os provérbios são expressões de uso generalizado, utilizados em diferentes situações conversacionais e assumindo diferentes funções no discurso em que se integram. Do ponto de vista sintático, este tipo de expressões apresenta uma grande variedade de estruturas. Tendo isto em conta, o objetivo principal deste trabalho é estabelecer uma classificação formal sintática dos provérbios do português europeu. Para tal, pretendemos desenvolver e aprofundar a tipologia de classificação formal proposta por Rassi *et al*. (2014). Por conseguinte, será considerada uma subclassificação para as classes que apresentam um elevado número de provérbios e uma eventual reclassificação de alguns dos tipos. A proposta de classificação foi validada pela anotação independente por dois linguistas de uma lista de provérbios muito usuais, medindo-se depois o acordo entre anotadores, que foi muito elevado. Esta classificação, por sua vez, será o ponto de partida para o desenvolvimento de um procedimento de classificação automática deste tipo de estruturas, e contribuir assim para a elaboração de recursos para diferentes aplicações em Processamento da Linguagem Natural (PLN).

Palavras-chave: Provérbios. Classificação sintática. Português europeu. PLN.

Proverbs are expressions of widespread use, appearing in different conversational situations, and assume different functions in the discourse in which they are integrated. From a syntactic point of view, this type of expression presents a great variety of structures. With this in mind, the main goal of this study is to determine a formal syntactic classification of the European Portuguese proverbs. For this, we intend to develop and extend the typology of the formal classification proposed by Rassi *et al*. (2014). Therefore, a subclassification for the classes with a large number of proverbs, and a possible reclassification of some of the types will be considered. The classification proposal was validated by having two linguists independently annotating a list of very usual proverbs, and then calculating the inter-annotator agreement, which was found to be very high. This classification, in turn, will be the starting point for the development of an automatic classification procedure, and thus contribute to the preparation of resources for different applications in Natural Language Processing (NLP).

Keywords: Proverbs. Syntactic classification. European Portuguese. NLP.

[*] Universidade do Algarve – FCHS, Campus de Gambelas, Faro, Portugal. 0000-0001-7709-6889:
[**] Universidade do Algarve – FCHS, Campus de Gambelas, Portugal | INESC-ID Lisboa – L2F, Lisboa, Portugal. 0000-0003-4603-4364:

•

1. Introdução

Os provérbios estabelecem relações discursivas de natureza diferente com o texto à sua volta. Desde os variados mecanismos utilizados para a sua introdução em discurso, mas mesmo na ausência destes, os provérbios são sentidos (e interpretados) como texto alheio, são microtextos citados, trazidos ao discurso a partir do conhecimento partilhado que os falantes têm do mundo e que se encontra cristalizado numa expressão linguística mais ou menos estável.

O emprego deste tipo de estruturas levanta diversos problemas à análise automática de textos. Na medida em que estes surgem como um texto encaixado noutro texto, os processos de coesão discursiva, como as relações de correferência ou o encadeamento de tempos-modos verbais são, muitas vezes, suspensos. Frequentemente, o seu ponto de inserção não é explicitamente assinalado, havendo situações de reutilização/variação criativa dos seus elementos lexicais mais característicos, que implicam o conhecimento prévio das formas de base para a sua completa descodificação. Assim, os provérbios constituem um desafio ao processamento da linguagem natural (PLN), exigindo uma identificação e delimitação mais precisas.

Pelo facto de o uso dos provérbios predominar na oralidade – o que está, talvez, na origem de certas características formais das expressões proverbiais, tais como as marcas mnemónicas de métrica, rima interna, paralelismo, etc.; um dos aspetos mais relevantes para a sua identificação é o tratamento da variação formal (sobretudo lexical) e a união das variantes sob uma única estrutura de base. Neste sentido, já foi desenvolvido um método que permite identificar automaticamente este tipo de expressões em textos, e que consiste na construção manual de transdutores que representam cada unidade paremiológica (provérbios + variantes) (Reis & Baptista 2016a, 2017b, 2017c) e que teve como ponto de partida o trabalho desenvolvido por Rassi *et al*. (2014) para o português do Brasil, que será descrito no ponto 2 deste trabalho, mas que deste difere não só pelo seu escopo/abrangência como também pela metodologia utilizada, uma vez que, se chegou à conclusão de que uma abordagem de geração exclusivamente automática não é suficientemente precisa para os objetivos que se pretendia alcançar.

O passo seguinte será o desenvolvimento de um procedimento de classificação automática deste tipo de estruturas. Esta classificação permitirá reunir expressões formalmente semelhantes de forma a poder aplicar sobre elas métodos de processamento de língua natural em diversas aplicações (*e.g.* identificação de variantes, reconhecimento de usos 'criativos' de provérbios, tradução, entre outros).

Esta classificação será desenvolvida sob o ponto de vista sintático, e partindo da classificação sintática proposta por Rassi *et al*. (2014) para os provérbios do português do Brasil. Tendo-se este propósito em mente, pretende-se agora estabelecer uma classificação formal sintática dos provérbios do português europeu, de forma a que, com base nesta classificação, seja posteriormente desenvolvido um procedimento de

classificação automática de uma base de dados de provérbios. O objetivo de tal classificação é permitir o desenvolvimento de métodos mais eficientes de reconhecimento automático de expressões proverbiais em textos e a sua aplicação em diferentes tarefas de PLN (Processamento da Linguagem Natural).

Este trabalho está estruturado da seguinte forma: a secção 2 apresenta os trabalhos relacionados; a secção 3 expõe e descreve as classes formais dos provérbios do português europeu; segue-se a secção 4 onde é apresentada a validação da taxonomia e os resultados obtidos; finalmente a secção 5 apresenta as conclusões e perspetivas de trabalhos futuros.

2. Trabalhos relacionados

Já existem várias propostas de classificação sintática das frases fixas do português europeu, nomeadamente alguns trabalhos sobre provérbios, porém, não há ainda um trabalho que apresente de forma sistemática uma tipologia deste tipo de expressões. Dos vários estudos sobre provérbios do Português Europeu (PE), poucos são que tratam esses provérbios sobre o ponto de vista sintático. L. Chacoto tem vindo a desenvolver trabalhos neste âmbito, fazendo uma análise léxico-sintática deste tipo de estruturas. Por exemplo, Chacoto (1994) analisa a fixidez e a variação dos provérbios e constata que estas estruturas, apesar da fixidez das suas formas, podem apresentar variação de diferentes tipos e resultante de diferentes fatores. Chacoto (2007) faz, também, um estudo comparativo dos provérbios do português europeu e do espanhol iniciados por *quem*/*quien*; em Chacoto (2008), a autora trata as estruturas comparativas dos provérbios do PE e, em Chacoto (2010), as estruturas condicionais.

Como já referimos num trabalho anterior (Reis & Baptista 2016b), para o português do Brasil, por sua vez, já foi apresentada uma proposta de classificação formal (sintática) de provérbios (Rassi *et al*. 2014), com base numa coleção de três mil quinhentos e dois provérbios (e suas variantes), organizados em quinhentos e noventa e quatro tipos ou formas de base e recolhidos de vários dicionários de provérbios do português do Brasil. Os autores apresentaram um método, baseado em máquinas de estados finitos (Clark, Fox & Lappin 2010), para identificação automática de provérbios em *corpora* de grandes dimensões, que foi testado no *corpus PLN-Br* constituído por Bruckschein *et al*. (2008), com vinte e nove milhões de *tokens* de texto jornalístico retirado da edição do jornal brasileiro *Folha de São Paulo* (1994–2005). Os autores reportaram uma precisão de 60% a 73%, dependendo sobretudo da classe do provérbio e do grau de completude das variantes lexicais registadas no léxico. O número de provérbios encontrados (e suas variantes), face aos de que o léxico dispunha, é, no entanto, reduzido, o que deverá estar diretamente relacionado com a natureza do *corpus*, uma vez que se trata de texto jornalístico.

Nesse trabalho (Rassi *et al*. 2014), os provérbios são classificados, tendo em conta o número de orações/proposições (ou frases elementares) que apresentam: uma, duas ou três orações, correspondendo às classes principais P1, P2 e P3, respetivamente; e também com base nos seus elementos-chave. De forma muito resumida, esta classificação consiste em:

P1
– Construções impessoais (*e.g. Não há regra sem exceção*) – P1F1;
– Construções com verbo copulativo (*e.g. O amor é cego*) – P1F2;
– Negação obrigatória (*e.g. Burro velho não aprende línguas*) – P1F4;
– Anteposição obrigatória do complemento preposicional (*e.g. Para bom entendedor meia palavra basta*) – P1F5;
– Outros provérbios com uma única oração (*e.g. A união faz a força*) – P1F3.

P2
– Orações comparativas (*e.g. Mais vale pouco que nada*) – P2F1;
– Orações coordenadas (*e.g. A palavra é de prata e o silêncio é de ouro*) – P2F2;
– Proposições coordenadas sem verbo (*e.g. Tal pai, tal filho*) – P2F3;
– Interrogativas-sujeito (*e.g. Quem ri por último, ri melhor*) – P2F4;
– Orações subordinadas (*e.g. Os amigos são muitos quando é grande a abastança*) – P2F5;
– Anteposição obrigatória da oração subordinada (*e.g. Quando um burro fala, os outros baixam as orelhas*) – P2F6.

P3
– Tripla coordenação (*e.g. A laranja de manhã é ouro, à tarde é prata e à noite mata*).

Neste trabalho, partindo da tipologia proposta pelos autores acima citados, elaborar-se-á uma tipologia formal para os provérbios do PE, que descreveremos no ponto seguinte deste trabalho, secção 3. Pretende-se com esta proposta, orientar o desenvolvimento de um procedimento de classificação automática dos provérbios de uma base de dados com cerca de cento e catorze mil expressões proverbiais (Reis & Baptista 2016a), coligidas a partir de 4 coletâneas de provérbios portugueses (Costa 1999; Machado 1996; Moreira 1996; Parente 2005).

3. Classificação formal (sintática) das expressões proverbiais do português europeu

Como já indicado anteriormente, o objetivo que norteia este estudo é o estabelecimento de um quadro de classificação formal (sintática) das expressões proverbiais em português europeu. Neste sentido, pretende-se desenvolver e aprofundar a tipologia de classificação formal, de natureza sintática, proposta por Rassi *et al.*, aplicada a provérbios usuais no português do Brasil, mas agora com uma maior base empírica e revendo alguns critérios ali usados, por questões de coerência/consistência interna e simplificação do procedimento de classificação.

Algumas das dificuldades desta classificação prendem-se com o facto de muitas destas expressões apresentarem características que podem ser associadas a diferentes classes formais, segundo a hierarquia taxonómica que as estrutura. De forma a ultrapassar este problema, ter-se-á por base, neste trabalho, uma classificação assente em critérios estritamente formais, o que auxiliará no processo taxonómico. Efetivamente, o critério de

base para a classificação de Rassi *et al*. (2014), nomeadamente, o número de proposições constitutivas dos provérbios, não foi o critério predominante para a classificação aqui apresentada, por se tratar de um critério que revelou uma reprodutibilidade problemática, já que corresponde a um construto teórico e não a uma propriedade formal, empiricamente evidenciada. Por outro lado, outros critérios como a negação obrigatória não foram considerados num primeiro momento, já que esta propriedade é transversal a frases de tipologias muito diversas.

A classificação sintática dos provérbios do português europeu está dividida em 8 classes de base, tendo algumas destas sido subdivididas, conforme mostraremos adiante. Saliente-se que esta subclassificação é um processo em curso, que na sequência da determinação da extensão dos conjuntos assim constituídos, poderá vir a sofrer atualizações/adaptações. De seguida, apresentamos as classes de base, já estabelecidas:

Tabela 1. Classificação formal dos provérbios do português europeu.

Classe	Estrutura	Definição	Exemplo
P1	*0 V w*	construções impessoais	Há males que vêm por bem.
P2	*V=Vcop*	construções predicativas	O amor é cego.
P3	N_1 N_2	construções sem verbo	Cada cavadela, sua minhoca.
P4	*QueF*	orações substantivas	Quem escorrega também cai.
P5	F_1 *Conj(subord)* F_2	orações subordinadas (adverbiais e adjetivas)	Quando a esmola é muita, o pobre desconfia.
P6	F_1 *Conj(subord-comp)* F_2	orações comparativa	Debaixo da manta, tanto vale a escura como a branca.
P7	F_1 *Conj(coord)* F_2	orações coordenadas	A conversa não engasga mas empata.
P8	N_0 *V w*	frases simples	Em abril, queima a velha o carro e o carril.

Fonte: elaborado pelos autores.

Legenda: ***N₀, N₁, N₂***: grupos nominais com a função, respetivamente, de sujeito, 1º complemento ou 2º complemento; ***0***: construção impessoal (sem sujeito); ***w***: sequência não especificada de elementos; ***V***: verbo; ***Vcop***: verbo copulativo; ***Conj***: conjunção (***subord***: subordinativa, ***subord-comp***: subordinativa comparativa, ***coord***: coordenativa); ***QueF***: oração subordinada substantiva; ***F₁, F₂***: frase ou oração.

P1: Construções impessoais – Esta classe é composta por construções sem sujeito em que os verbos são conjugados apenas na 3ª pessoa do singular. Constam desta classe o verbo *haver* (na aceção de *existir*), por exemplo, *Há males que vêm por bem* e os verbos que exprimem fenómenos atmosféricos e da natureza, por exemplo, *Em Março chove cada dia um pedaço*.

A proposta de Rassi *et al*. (2014) contempla igualmente uma classe de construções impessoais, que difere da que propomos na medida em que abrange as construções com o verbo *ter* (impessoal) conjugado apenas na 3ª pessoa do singular (na aceção de *existir*, *e.g. Tem muito cacique pra pouco índio*, sendo sinónimo de *haver*, *e.g. Há muito cacique pra pouco índio*), usadas na variante do português do Brasil; e abrange ainda construções com sujeito indefinido, marcadas pelo pronome indefinido *se*, por exemplo, *Devagar se vai ao longe*, que aqui foram integradas noutras classes.

P2: Construções predicativas – Esta classe é constituída por construções que constituem frases simples (com um único predicado) e que apresentam um verbo copulativo (*e.g. ser*, *estar*, *ficar*, *continuar*, *parecer*, *permanecer*, *tornar-se*, *revelar-se*, etc.) e ligam o sujeito a um outro constituinte, que pode ter como núcleo:

i) um grupo nominal [P2-1], por exemplo, *Dinheiro é* ***remédio***;
ii) um grupo adjetival [P2-2], por exemplo, *Amor de parente é mais* ***quente***;
iii) um grupo preposicional [P2-3], por exemplo, *Amigo de mesa não é* ***de firmeza***;
iv) ou um grupo adverbial [P2-4], por exemplo, *A boa educação fica* ***bem*** *em todo o lado*.

Não foram incluídas nesta classe, por corresponderem à definição de outras classes, as seguintes estruturas/construções:

a) as estruturas comparativas que apresentam verbos copulativos, por exemplo, *Cada um é* ***como*** *cada qual* ou ***Melhor é*** *dobrar* ***que*** *quebrar*; estas integrarão a classe P6;
b) as orações subordinadas substantivas interrogativas indiretas, por exemplo, *Feliz é* ***quem feliz se julga*** ou *O pior cego é* ***o que não quer ver***; estas integrarão a classe P4;
c) as orações subordinadas infinitivas, por exemplo, *Partir é* ***morrer*** *um pouco*; estas integrarão a classe P4;
d) as orações reduzidas (substantivas) que equivalem a um grupo nominal com função de sujeito, por exemplo, ***Errar*** *é humano*; estas integrarão a classe P4;
e) outras frases que integram orações subordinadas (adjetivas e adverbiais), por exemplo, *O sol quando nasce é para todos*; estas integrarão a classe P5;
f) outras frases que integram orações coordenadas, por exemplo, *O dinheiro é bom companheiro, mas mau conselheiro*); estas integrarão a classe P7;
g) e as construções que apresentam *é que* (muitas vezes considerado um elemento expletivo/enfático), e que não apresentam nenhuma das características já descritas nas alíneas anteriores, por exemplo, *A intenção é que conta* (*cf. Não é só a intenção que conta*); de um modo geral, a expressão *é que* não é considerada na classificação.

Foi necessário subdividir esta classe dado o elevado número de provérbios que apresentam um verbo copulativo na sua constituição.

Para o português do Brasil foi igualmente proposta uma classe que abrange as construções com verbo copulativo, a qual contempla os verbos *ser* e *estar*.[1]

P3: Construções sem verbo – Esta classe integra as construções que não apresentam verbo (construções elípticas), por exemplo, *Muita parra, pouca uva*. Equivale à classe P2F3 da classificação proposta por Rassi *et al*. (2014).

P4: Orações substantivas – Esta classe inclui as orações subordinadas substantivas completivas e relativas sem antecedente (interrogativas indiretas), por exemplo, *Quem*

[1] A classificação apresentada por Rassi *et al*. (2014) não elenca, porém, nenhum provérbio deste tipo com o verbo copulativo *estar*. Trata-se de expressões como, por exemplo, *No meio é que está a virtude*. Encontramos, além disso, provérbios com o verbo *estar* integrados em outras classes, nomeadamente os provérbios *O mal está nos olhos de quem o vê* e *As melhores essências estão nos menores frascos*, que constam da classe P1F3 (N_0 *v w*).

tudo quer tudo perde; *O que é doce nunca amargou* ou *Onde entra o sol, não entra o médico*, e equivale à classe P2F4 de (*idem*: *ibidem*), mas incluindo agora as orações introduzidas pelo advérbio *onde*, sem nome ou expressão a que este esteja associado (sem referente).

P5: Orações subordinadas (adverbiais e adjetivas) – Desta classe constam as orações subordinadas adverbiais (não comparativas) e as orações subordinadas adjetivas (relativas). É ainda tido em consideração se a locução ou conjunção que introduz estas orações se encontra anteposta à oração principal [P5-1], por exemplo, *Quando a esmola é muita o pobre desconfia*; ou depois da oração principal [P5-2], *Mal ladra o cão, quando ladra de medo*; o que determinou o desenvolvimento das subclasses [P5-1] e [P5-2]. No que respeita à subordinação, e considerando a ordem das orações (principal e subordinada), Rassi *et al.* (2014) consideraram esta como critério taxonómico, constituindo duas classes distintas de provérbios (P2F5 e P2F6), que aqui aparecem sob a mesma classe principal (P5).

P6: Orações comparativas – Esta classe integra as orações subordinadas adverbiais comparativas (introduzidas por conjunções ou locuções comparativas), por exemplo, *Voa o tempo como o vento*, *Mais vale tarde que nunca*, *Antes dobrar que quebrar* ou *Tal é o Demo como sua mãe*[2]; esta classe corresponde à classe P2F1 de Rassi *et al.* (2014).

P7: Orações coordenadas – Desta classe constam as orações coordenadas (com verbo explícito); estas podem ser sindéticas, por exemplo, *A justiça tarda mas não falta*; ou assindéticas (sem a conjunção coordenativa expressa); esta classe equivale à classe P2F3 da classificação proposta por Rassi *et al.* (2014).

P8: Frases simples – Esta classe é constituída por todas as frases simples que não tenham sido previamente integradas nas classes definidas anteriormente; estas expressões podem apresentar inversão dos constituintes [P8-1], por exemplo, *Em abril queima a velha o carro e o carril* ou podem atender à ordem básica dos constituintes [P8-2], por exemplo, *Bom madeiro corta-se em janeiro*. Para o português do Brasil foram estabelecidas duas classes distintas para representar estas frases com uma única proposição: P1F5 quando há anteposição de complemento e são iniciadas por preposição; e P1F3 quando apresentam os seus elementos na ordem básica.

De seguida iremos apresentar a metodologia utilizada para validar a taxonomia proposta.

[2] Ainda que alguns autores incluam nesta classe as orações que expressam conformidade e proporção (nomeadamente Chacoto (2008)), estas não foram aqui incluídas, já que constam da classe P5. Seguimos a proposta de classificação de Mateus *et al.* (2003, p. 762), que consideram que "Ao contrário das orações comparativas, as orações conformativas são deslocáveis (...), podem ser objeto de clivagem (...), e são adjuntos".

4. Validação da taxonomia: classificação do mínimo paremiológico do português e resultados

Com vista a determinar a reprodutibilidade dos critérios taxonómicos acima enunciados, procedeu-se à classificação sistemática do *mínimo paremiológico* do português europeu. Por *mínimo paremiológico* entende-se o conjunto de provérbios mais conhecidos e frequentemente utilizados de uma comunidade cultural (Permjakov 1973). Neste caso, trata-se de um conjunto de trezentos e dezassete expressões proverbiais, cuja disponibilidade lexical foi atestada de múltiplas formas (com recurso a informantes, dados de *corpora* e frequências observadas em textos na internet) (Reis & Baptista 2017a).

Para tal, dois anotadores, independentemente, procederam à classificação das expressões do *mínimo paremiológico*, utilizando para isso um conjunto inicial de diretivas de anotação em que se explicitavam os critérios e se davam exemplos e contraexemplos. Os anotadores foram dois linguistas, falantes nativos do português europeu, com conhecimento de todos os provérbios alistados, bem como das condições pragmáticas do seu uso.

Realizada a anotação, os dados foram comparados, utilizando-se o software *ReCal 0.1 Alpha para 2 anotadores*[3] para calcular a concordância entre anotadores. Das trezentas e dezassete expressões, os 2 anotadores concordaram trezentas e dez vezes (97,8%) e discordaram apenas 7 (Krippendorff alfa = 0.973). Tal representa um grau de concordância muito elevado.

Os casos em que houve discordância foram os seguintes (*vd.* Tabela 2):

Tabela 2. Classificações diferentes.

ID-Prov	Provérbio	Anot 1	Anot 2
MP_P47	Aprender até morrer.	P4	P5
MP_P70	Cada um sabe as linhas com que se cose.	P4	P5
MP_P111	Errar é humano.	P2	P4
MP_P112	Faz o que eu digo e não faças o que eu faço.	P7	P4
MP_P151	Não se pode ter tudo.	P4	P8
MP_P157	Ninguém é de ferro.	P2	P8
MP_P170	O fruto proibido é o mais apetecido.	P6	P2

Fonte: elaborado pelos autores.

Na expressão MP_47, houve confusão do primeiro anotador quanto ao estatuto da oração subordinada *até morrer*; em MP_70, *idem*, relativamente à oração adjetiva relativa *com que se cose*; em MP_111, *idem*, relativamente à oração reduzida (substantiva) *errar*; em MP_112, o primeiro anotador não considerou as orações interrogativas indiretas (critério num patamar hierarquicamente superior), dando prioridade à coordenação; em MP_151, o primeiro anotador, certamente por lapso, selecionou uma classe incorreta, já que *poder* é apenas um verbo auxiliar modal de *ter*, aqui empregue como verbo pleno; em MP_157, o segundo anotador não considerou o verbo copulativo; por último, em MP_170, o

[3] Disponível em: http://dfreelon.org/utils/recalfront/recal2/#doc.

primeiro anotador escolheu a classe P6, considerando a comparação (sem segundo termo de comparação) em detrimento do verbo copulativo (critério superior).

Após esta fase, as divergências foram resolvidas, e os critérios de classificação "afinados" de modo a serem mais consistentes e explícitos.

Verifica-se, portanto, que os critérios de classificação são suficientemente claros, explícitos e reprodutíveis para poderem ser empregues com segurança na tarefa de classificação das restantes expressões proverbiais da base de dados (com centro e catorze mil expressões) que constitui o nosso *corpus* de trabalho.

Esta listagem do *mínimo paremiológico* com a respetiva classificação sintática será o embrião que levará à constituição de uma coleção dourada (*golden standard*), que permita avaliar métodos de classificação automática de provérbios (trabalho em curso). Esses métodos deverão, então, permitir outros processamentos mais complexos, tais como a identificação de variantes de provérbios sob uma mesma entidade, a unidade paremiológica.

5. Conclusões e trabalho futuro

Neste trabalho foi apresentada uma proposta de classificação formal (sintática) para os provérbios do português europeu. Esta proposta desenvolve e aprofunda a classificação de Rassi *et al*. (2014), mas com uma maior base empírica, e revê alguns desses critérios, a fim de garantir uma maior consistência interna e uma simplificação do procedimento de classificação. Em concreto, abandonámos o critério daqueles autores quanto ao número de proposições, por se tratar de um construto teórico e não de um padrão formal, diretamente observável. Também o critério da negação obrigatória foi preterido nesta fase, já que é transversal a várias classes formais. Além disso, os critérios de classificação foram ordenados para uma aplicação sucessiva, a fim de permitir o desenvolvimento de um algoritmo de classificação, que permitisse no futuro automatizar o processo, usando métodos de processamento de língua natural.

Foram, assim, constituídas 8 classes formais principais, baseadas no tipo de construção sintática que o provérbio exibe, nomeadamente construções impessoais, predicativas (com verbo copulativo), sem verbo, subordinadas (substantivas), subordinadas (adverbiais e adjetivas), construções comparativas, orações coordenadas e outras frases simples.

A estrutura taxonómica constituída foi avaliada através da classificação sistemática do *mínimo paremiológico*, feita autonomamente por dois anotadores, revelando um grau de concordância bastante elevado (97,8%). As diferenças observadas permitiram melhorar as diretivas de classificação, tornando os critérios mais claros e mais explícitos.

Os critérios de classificação são suficientemente claros, explícitos e reprodutíveis para poderem ser empregues com segurança na tarefa de classificação das restantes expressões proverbiais da base de dados.

No futuro, pretendemos desenvolver procedimentos para a classificação automática de expressões proverbiais com base neste trabalho; e eventualmente métodos que venham complementar os que usamos neste momento para identificar expressões proverbiais em

textos, em especial as que não se encontram ainda recenseadas ou que resultem de uma variação criativa pelos falantes.

Financiamento:
Esta investigação foi parcialmente suportada por fundos nacionais através da Fundação para a Ciência e a Tecnologia (ref. UID/CEC/50021/2019).

Referências

Bruckschen, M., Muniz F., Souza, J. G. C., Fuchs, J. T., Infante, K., Muniz, M., ... Aluísio, S. (2008). Anotação linguística em XML do Corpus PLN-BR. *Série de Relatórios do NILC, NILC–NILC-TR-08-09*. Universidade de São Paulo, Brasil. Disponível em: http://www.nilc.icmc.usp.br/nilc/index.php/publications#conference_papers

Chacoto, L. (1994). *Estudo e formalização das propriedades léxico-sintácticas das expressões fixas proverbiais* (Tese de mestrado, Faculdade de Letras da Universidade de Lisboa, Lisboa).

Chacoto, L. (2007). A sintaxe dos proverbios. As estruturas quem/quien en português e español. *Cadernos de Fraseoloxía Galega*, *9*, 31–53.

Chacoto, L. (2008). Vale mais um gosto na vida que três vinténs na algibeira - Las estructuras comparativas en los proverbios portugueses. In G. Conde Tarrío (Ed.), *Aspectos formales y discursivos de las expresiones fijas* (pp. 87–103). Frankfurt: Peter Lang.

Chacoto, L. (2010). Não há rifão velho, se é dito a propósito – La condición en los refranes portugueses. In J. Korhonen, W. Mieder, E. Piirainen, & R. Piñel (Eds.), *Actas do Congresso Internacional Europhras 2008* (pp. 58–65). Universidade de Helsínquia, Finlândia.

Clark, A., Fox, C. & Lappin, S. (Eds.) (2010). *The handbook of computational linguistics and natural language processing*. New Jersey: Wiley-Blackwell.

Costa, J. (1999). *O Livro dos Provérbios Portugueses*. Lisboa, Portugal: Editorial Presença.

Machado, J. (1996). *O Grande Livro dos Provérbios* (1ª ed.). Lisboa: Editorial Notícias.

Mateus, M. H. M., Brito, A. M., Duarte, I., Faria, I. H., Frota, S., Matos, G., … Villalva, A. (2003). *Gramática da Língua Portuguesa* (5ª ed. revista e aumentada). Lisboa: Caminho.

Moreira, A. (1996). *Provérbios Portugueses*. Lisboa: Editorial Notícias.

Parente, S. (2005). *O Livro dos Provérbios* (1ª ed.). Lisboa: Editora Âncora.

Permjakov, G. L. (1973). On the paremiological level and premiological minimum of language. *Proverbium*, *22*, 862–863.

Rassi, A., Baptista, J. & Vale, O. (2014). Automatic detection of proverbs and their variants. In M. J. Pereira, J. P. Leal & A. Simões (Eds.), *3rd Symposium on Languages, Applications and Technologies (SLATE'14)*, (pp. 235–249). Leibniz: Dagstuhl Publishing. https://doi.org/ 10.4230/OASIcs.SLATE.2014.235

Reis, S. & Baptista, J. (2016a). Let's play with proverbs? – NLP tools and resources for iCALL applications around proverbs for PFL. In *Proceedings of the International Congress on Interdisciplinarity in Social and Human Sciences*, 5-6 maio (pp. 427–446). Faro: Universidade do Algarve. https://doi.org/10.13140/RG.2.1.1244.7603

Reis, S. & Baptista, J. (2016b). Portuguese Proverbs: Types and Variants. In G. C. Pastor (Ed.), *Computorised and corpus-based approaches to phraseology: Monolingual and multilingual perspectives* (pp. 208–2017). Geneve: Editions Tradulex.

Reis, S. & Baptista, J. (2017a). Estimating lexical availability of European Portuguese proverbs. In R. Mitkov (Ed.), *Lecture Notes in Computer Science: Vol. 10596. Computational and corpus-based phraseology. EUROPHRAS 2017* (pp. 232–244). Cham: Springer. https://doi.org/10.1007/978-3-319-69805-2_17

Reis, S. & Baptista, J. (2017b). O uso de provérbios no ensino de português. In R. Soares & O. Lauhakangas (Eds.), *10º Colóquio interdisciplinar sobre provérbios. Actas ICP16* (pp. 521–538). Tavira, Portugal: AIP–IAP.

Reis, S. & Baptista, J. (2017c). Os provérbios em manuais de ensino de Português Língua Não Materna. In V. Pinheiro & G. H. Paetzold (Eds.), *Jornadas de Descrição do Português, integradas no STIL2017*, 2–5 outubro (pp. 247–255). Uberlândia: Sociedade Brasileira de Computação.

[recebido em 1 de junho de 2019 e aceite para publicação em 19 de outubro de 2019]