Acta Scientiarum. Language and Culture
ISSN: 1983-4675
ISSN: 1983-4683
actalan@uem.br
Universidade Estadual de Maringá
Brasil

Determinação de um mínimo paremiológico do português europeu

Reis, Sónia; Baptista, Jorge
Determinação de um mínimo paremiológico do português europeu
Acta Scientiarum. Language and Culture, vol. 42, núm. 2, e52114, 2020
Universidade Estadual de Maringá, Brasil
Disponible en: https://www.redalyc.org/articulo.oa?id=307466046014
DOI: https://doi.org/10.4025/actascilangcult.v42i2.52114

linguística

Determinação de um mínimo paremiológico do português europeu

Establishing the paremiological minimum of European Portuguese

Sónia Reis
Universidade do Algarve, Portugal
reis.soniamm@gmail.com

Jorge Baptista
Universidade do Algarve, Portugal
Instituto de Engenharia de Sistemas e Computadores, Investigação e Desenvolvimento de Lisboa, Portugal

DOI: https://doi.org/10.4025/actascilangcult.v42i2.52114
Redalyc: https://www.redalyc.org/articulo.oa?id=307466046014

Recepción: 04 Febrero 2020
Aprobación: 04 Junio 2020

Resumo:

O principal objetivo deste estudo é apresentar o 'mínimo paremiológico' do português europeu, isto é, a lista dos provérbios mais conhecidos e mais frequentemente utilizados pela generalidade dos falantes da comunidade linguística do português europeu. Para o estabelecimento do 'mínimo paremiológico' do português europeu foram utilizados diferentes procedimentos metodológicos ao longo de cinco anos. Numa primeira fase, coligiu-se uma base de dados digital com mais de 114.000 entradas (provérbios e variantes) e fez-se uma primeira seleção manual dos provérbios considerados mais usuais, tendo-se depois calculado a concordância entre anotadores, que foi bastante elevada. Procedeu-se também ao cálculo da frequência de provérbios e variantes (n. de ocorrências) em várias fontes: (1) em dicionários e coletâneas de provérbios; (2) num *corpus* de textos jornalísticos (*CETEMPúblico*, Santos & Rocha, 2001); (3) em manuais escolares de Português e de Português Língua não Materna; e (4) em dois motores de busca (Google e Bing). Procedeu-se ainda à aplicação de dois questionários distintos, *online*. Obteve-se assim uma lista de 318 provérbios – o mínimo paremiológico do português europeu. A lista dos 318 provérbios poderá# ter diversas aplicações, quer para o desenvolvimento de instrumentos de diagnóstico ou terapia de certas patologias da linguagem, quer para a aprendizagem de português como língua estrangeira.

Palavras-chave: disponibilidade lexical, linguística de *corpus* , provérbio, mínimo paremiológico, português europeu, variação.

Abstract:

The main objective of this study is to present the paremiological minimum of European Portuguese, i.e., the list of the most widely known and frequently used proverbs employed by most speakers of the European Portuguese language community. To establish the paremiological minimum of European Portuguese, different methodological procedures have been applied, over a period of five years. In the first phase, a digital database with more than 114,000 entries (proverbs and variants) was collected and a first manual selection of the most common proverbs was made, and then the agreement between annotators was calculated, which was quite high. The frequency of proverbs and variants (number of occurrences) was also calculated in several sources: (1) in dictionaries and collections of proverbs; (2) in a corpus of journalistic texts (*CETEMPúblico*, Santos & Rocha, 2001); (3) in Portuguese and Portuguese as a Foreign Language textbooks; and (4) in two search engines (Google and Bing). Two distinct online questionnaires were also applied. A list of 318 proverbs – the paremiological minimum of European Portuguese - was thus obtained. The list of the 318 proverbs may have several applications, either for the development of diagnostic or therapeutic tools for certain language pathologies, or for learning Portuguese as a foreign language.

Keywords: lexical availability, *corpus* linguistics, proverbs, paremiological minimum, European Portuguese, variation.

Notas de autor

reis.soniamm@gmail.com

Introdução

O interesse em determinar um conjunto de provérbios comum, conhecido pela comunidade linguística que os utiliza no dia a dia, tem sido partilhado entre profissionais de diferentes áreas do conhecimento.

Mieder (1992, 1994) reflete sobre investigações desenvolvidas por psicólogos e outros estudiosos, no âmbito de determinar o conhecimento que as pessoas têm de provérbios, e explica quais os provérbios particularmente bem conhecidos nos Estados Unidos. O autor acrescenta que estes provérbios fazem parte da cultura dos falantes do inglês, e que os mais comuns destes formam um mínimo paremiológico, usando as próprias palavras do autor, 'minimum of proverbial knowledge', que é preciso conhecer de forma a se poder comunicar efetivamente em inglês. Acrescenta ainda que este assunto é de grande importância para os lexicógrafos que estão envolvidos na escrita de dicionários de língua estrangeira, ou mesmo para os professores que ensinam inglês. Para o autor, a literacia e a cultura paremiológica andam de mãos dadas.

Permjakov (1973) foi, que saibamos, um pioneiro no sentido de ser o primeiro a estabelecer o conjunto de provérbios mais conhecidos e frequentemente utilizados de uma comunidade cultural, mais especificamente da população de Moscovo. O autor determinou um mínimo paremiológico russo, que incluía cerca de 300 provérbios.

Outros estudos têm vindo a ser desenvolvidos neste âmbito, em várias línguas, designadamente a série 'Mínimo paremiológico', publicada na Biblioteca paremiológica y fraseológica (Ruiz-Ayu#car & Muñoz, 2016), do Centro Virtual Cervantes. Esta série inclui um conjunto de estudos sobre 'paremias'. O seu objetivo principal é observar o uso dessas 'paremias' do espanhol (de Espanha) na atualidade e determinar quais destas é que devem estar presentes no processo de ensino-aprendizagem do espanhol. Um outro objetivo do projeto é estabelecer uma 'correspondencia' entre esse mínimo paremiológico do espanhol e outras línguas de trabalho, nomeadamente o francês, o alemão, o italiano, o português, entre outras. Com este termo 'correspondencia' pretende-se designar a relação entre fraseologismos de diferentes línguas, isto é, a equivalência entre fraseologismos de uma língua de partida e uma língua de chegada, de modo a evitar a possível perda de informação, muitas vezes associada às traduções literais das expressões deste género (Ruiz-Ayu#car & Muñoz, 2016). No quadro desse projeto, tem vindo a desenvolver-se um Mínimo Paremiológico do Português (ainda inédito), tendo como objetivo geral determinar o número de 'paremias' que fazem parte da competência linguística dos falantes nativos dos países que têm o português como língua oficial. Esse Mínimo do Português Europeu (Portugal continental e insular) será composto por 329 'paremias' e será publicado [1] na série 'Mínimo paremiológico'.

Foi com base na metodologia utilizada para determinar o mínimo paremiológico do espanhol que foram desenvolvidos os estudos análogos para a determinação dos 'minima' paremiológicos das restantes línguas que fazem parte do mesmo projeto. Os autores utilizaram diferentes estratégias para poder estabelecer um repertório paremiológico mínimo, por exemplo, solicitando aos informantes que lhes indicassem espontaneamente quais as 'paremias' populares que conhecem; ou apresentando-lhes uma lista da qual estes deveriam indicar quais as 'paremias' que conhecem e se as usam; ou ainda apresentando-lhes um conjunto de 'paremias' truncadas, mostrando apenas o seu primeiro membro, com o objetivo de os falantes as completarem. Após delimitarem o mínimo paremiológico, os autores verificaram em diferentes fontes documentais (imprensa, literatura e ensaios) e em contextos orais (meios de comunicação e informantes) se essas 'paremias' eram efetivamente utilizadas pelos falantes.

Gostaríamos, no entanto, de ter visto descritos os procedimentos para a determinação dos 'minima' paremiológicos das demais línguas que fazem parte daquele projeto (especificamente para o caso do português), com os quais já foram estabelecidas 'correspondencias' com as 'paremias' do espanhol.

No Refranero multilingue [2] encontram-se 1.601 provérbios portugueses. Na ficha técnica destes provérbios é indicado se cada um destes é 'De uso actual'; 'Muy usado'; 'Poco usado' ou 'En desuso'. Não

há, contudo, qualquer informação sobre como foi calculada esta informação. Foi esta uma das lacunas no conhecimento paremiológico do português que o nosso estudo visa preencher.

Diferentes métodos têm vindo a ser utilizados para a determinação dos 'minima' paremiológicos de diferentes comunidades culturais. Falaremos de alguns desses trabalhos já a seguir.

Ďurčo (2015) descreve as diferentes pesquisas que desenvolveu, de modo a estabelecer um 'Paremiological Optimum' para o eslovaco (Ďurčo, 2004, 2005a, 2005b, 2006, 2007).

O autor, numa primeira fase, procurou determinar quais das unidades paremiológicas registadas nas recolhas lexicográficas eram familiares aos falantes de eslovaco e eram por estes usadas. Considerando o elevado número de provérbios e variantes que se encontram nestas recolhas, o autor procurou limitar o corpus de trabalho. Para tal, foram necessários 5 passos: (i) a análise do registo lexicográfico de provérbios em dicionários antigos e atuais; (ii) a determinação de um conjunto principal de provérbios, após a redução do material paremiológico, feita por diferentes especialistas (linguistas, lexicógrafos, paremiólogos e etnólogos); (iii) a aplicação de um questionário on-line com o intuito de verificar se os inquiridos conheciam os provérbios selecionados pelos especialistas e se tinham familiaridade com esses provérbios; (iv) a análise das ocorrências em 'corpora' dos provérbios e suas variantes selecionados pelos inquiridos no questionário; e (v) a comparação e correlação dos provérbios considerados mais familiares pelos inquiridos com os provérbios mais frequentes em 'corpora'.

O questionário desenvolvido por este autor apresentava 2.834 provérbios e deveria ser descarregado pelos usuários, devidamente preenchido, e depois enviado via internet. O questionário apresentava uma parte inicial destinada a recolher a informação pessoal de cada respondente, nomeadamente o sexo, a idade, a escolaridade, as regiões onde viveu e a região onde vivia. Depois de preenchidos os dados pessoais, o respondente deveria escolher uma palavra-passe e, só então, ser-lhe-iam apresentados os provérbios, um de cada vez, e 4 opções de resposta: (i) 'Conheço e uso o provérbio'; (ii) 'Conheço, mas não uso o provérbio'; (iii) 'Não conheço o provérbio, mas compreendo-o' [3]; (iv) 'Não conheço o provérbio e não o compreendo'. Uma quinta opção também era apresentada, na qual o respondente poderia colocar uma variante que conhecesse desse provérbio. O programa permitia ainda que fossem alteradas as respostas dadas e que fossem adicionados, no final do questionário, outros provérbios que o respondente conhecesse e usasse, e que não tivessem sido contemplados pelo questionário. Uma outra característica deste questionário era permitir que o respondente interrompesse o seu preenchimento sempre que pretendesse.

O questionário obteve 42 respostas. Os respondentes, 24 mulheres e 18 homens, eram de diferentes regiões da Eslováquia e apresentavam uma idade média de 43 anos.

Apenas um pequeno conjunto de provérbios era conhecido e usado pelos respondentes (cerca de 16%, incluindo diferentes variantes). Com base nas respostas, foram formadas 5 categorias de provérbios: (i) o primeiro conjunto representa os provérbios que faziam parte do vocabulário ativo dos respondentes, e incluía aproximadamente 250 provérbios (uso declarado por mais de 50% dos entrevistados); (ii) um outro conjunto representa os provérbios que eram familiares aos respondentes, mas que não faziam parte do seu vocabulário ativo, e englobava 50 provérbios (em conformidade com mais de 50% dos inquiridos); (iii) o maior conjunto corresponde aos provérbios que 'não' eram conhecidos pelos respondentes, mas que eram por estes interpretados, e apresentava cerca de 1.900 provérbios; (iv) 100 dos 2.834 foram selecionados pelos respondentes como 'não conhecidos' e 'não interpretáveis' (em conformidade com mais de 50% dos inquiridos); e, finalmente, (v) apenas 2% dos respondentes apresentou outras variantes de provérbios que conhecia. Com base nos resultados obtidos foram então selecionados os provérbios que foram integrar o mínimo paremiológico.

De seguida, o conjunto de provérbios que fazem parte do vocabulário ativo dos respondentes foi comparado com a sua frequência de ocorrência no Slovak National Corpus [4]. Os resultados obtidos demonstram que 4 dos provérbios selecionados pelos respondentes não apresentavam qualquer ocorrência no corpus; 11 dos provérbios que foram considerados familiares pelos respondentes apresentavam um baixo

número de ocorrências; e apenas 17 dos provérbios selecionados apresentam mais de 50 ocorrências no corpus em análise. A correlação entre, por um lado, o nível de conhecimento/familiaridade com os provérbios indicados pelos inquiridos e, por outro lado, os valores de frequência dos provérbios no corpus permitiu ao autor estabelecer o mínimo paremiológico da língua eslovaca.

O mesmo método foi usado também para a língua eslovena (Meterc, 2012, 2014), o que permitiu comparar formal, semântica e pragmaticamente ('diasystemic') a equivalência entre provérbios eslovacos e eslovenos, selecionados com base numa mesma metodologia empírica (Ďurčo & Meterc, 2014).

Em suma, o mínimo paremiológico de uma língua deverá não só conter os provérbios que são parte de um fundo comum dessa língua, como também aqueles que são efetivamente utilizados pelos falantes.

Contudo, falar de provérbios é falar de variação (Chacoto, 1994), o que implica, em contraponto, a noção de que, intersetando as múltiplas variantes de um provérbio está uma unidade conceptual e pragmática, uma 'invariante', a que chamamos 'unidade paremiológica' (UP). Ainda que a expressão 'unidade paremiológica' não seja completamente inusitada - veja-se por exemplo a discussão terminológica de Gasanova e Taibova (2016), que compara 'saying', 'proverb' e o termo 'paroemi (a paremiological unit)' - acreditamos que o conceito de UP que aqui definimos está subjacente, de forma mais ou menos implícita, em numerosos trabalhos da literatura. No entanto, a sua utilização aqui implica, quanto a nós, um esforço na determinação dos critérios (formais, semânticos e pragmáticos) que permitem considerar diferentes formas como variantes de uma mesma UP (Reis & Baptista, 2016c).

No estabelecimento do mínimo paremiológico (MP) deverá ter#se em conta, pois, o fenómeno da variação, já que, de entre as variantes associadas à mesma UP, provavelmente uma é mais usual (eventualmente mais do que uma).

O MP deve ainda ser considerado um objeto dinâmico, uma lista que deverá ser atualizada ao longo do tempo, de modo a representar o conhecimento linguístico de uma dada comunidade, num dado momento.

Seguidamente, apresentaremos a base de dados digital de provérbios que serviu como ponto de partida para o estabelecimento do mínimo paremiológico do português europeu por nós estabelecido.

Para a constituição de uma base de dados digital de provérbios (Reis & Baptista, 2016a), procedeu#se à digitalização de quatro obras de cariz paremiológico (dicionários e coletâneas; Costa, 1999, Machado, 1996, Moreira, 1996, Parente, 2005). Numa primeira fase, atribuiu#se um código convencional a cada uma das expressões, de forma a que fossem identificadas univocamente consoante a fonte de onde foram recolhidas; depois, foi necessário efetuar um conjunto de transformações na listagem inicial, já que continha erros procedentes da digitalização; foi ainda necessário uniformizar essa listagem, de forma a que a pudéssemos tratar com ferramentas de processamento de texto - retirando espaços em branco e sinais de pontuação espúrios, retirando as notas e informações contidas no texto original, e outras informações adicionais, e.g. GLP_JPM05358 "Cansa quem dá, não cansa quem toma '(XVII, Delic, p. 189)'" (e.g. Machado, 1996, p. 134); LP_SP13673 "Em S. Lourenço '(10/8)', vai à vinha e enche o lenço" (e.g. Parente, 2005, p. 247); LP_SP04377 "Ano de salmões, ano de paixões '(Minho)'" (e.g. Parente, 2005, p. 84). Numa fase posterior, iniciámos outra tarefa, que consistiu em assinalar todas as expressões que não são provérbios, (nomes compostos, adjetivos compostos, frases fixas, ou outras). Neste momento, já assinalámos cerca de 4.600 destas expressões. Dada a dimensão da base de dados - mais de 114.000 entradas - esta tarefa ainda se encontra em curso.

Paralelamente a esta tarefa, estão também a ser desdobradas algumas entradas que apresentam variantes e.g. LP_SP04322 "Animal de bico nunca fez o 'dono (amo)' rico" (e.g. Parente, 2005, p. 84); GLP_JPM05553 "Casa de pai, vinha de avô. Var.: Casa de pai, vinha de avó" (e.g. Machado, 1996, p. 138). Presentemente, já desdobrámos 277 entradas na nossa base de dados e temos cerca de 5.600 entradas assinaladas para desdobrarmos.

A base de dados de provérbios assim constituída foi o ponto de partida para a determinação do mínimo paremiológico do português europeu.

Determinando um mínimo paremiológico para o português europeu

Da interseção dos vários métodos, constituímos o mínimo paremiológico do português europeu (MP), que inclui 318 provérbios. Trata-se de uma lista dinâmica, passível de ser modificada/ajustada, em função de novos dados que possam surgir. Seguidamente, descreveremos, de modo sucinto, cada uma das etapas que nos permitiram determinar a lista dos provérbios muito usuais.

Após a constituição da base de dados digital de provérbios, procedeu-se à inventariação das expressões mais frequentes (Reis & Baptista, 2016c). A frequência dos provérbios na base de dados pareceu-nos ser um indicador dos provérbios mais usuais, uma vez que foi constituída a partir de diferentes recolhas e dicionários de diferentes autores.

Para podermos agrupar esses provérbios, procurámos identificar as palavras#chave comuns entre eles, de forma semiautomática e com base em ferramentas de processamento da linguagem natural. Por palavras#chave entendemos os elementos lexicais plenamente significativos. Por exemplo, o provérbio 'Antes tarde do que nunca' tem como palavras#chave ['tarde nunca']. Para esse fim, removemos todas as stopwords ou seja, as palavras gramaticais que, de um modo geral, não contribuem para a identificação dos provérbios (sobretudo, determinantes, pronomes, conjunções e preposições).

Verificou-se, contudo, que as variantes de um mesmo provérbio nem sempre apresentam as mesmas palavras-chave, resultado do processo automático de construção dessas mesmas chaves, como acontece em:

(1a) Antes tarde que nunca [tarde nunca] (e.g. Parente, 2005, p. 90).
(1b) Mais vale tarde do que nunca [mais vale tarde nunca] (e.g. Parente, 2005, p. 333).
(1c) Melhor é tarde que nunca [melhor é tarde nunca] (e.g. Parente, 2005, p. 345).
(1d) Vale mais tarde do que nunca [vale mais tarde nunca] (e.g. Parente, 2005, p. 730).

Há, também, casos em que os provérbios apresentam as mesmas palavras#chave, mas que estas correspondem a provérbios diferentes, o que tem de ser analisado manualmente e caso a caso. Veja-se, a propósito, os exemplos (2) e (3) cuja chave é ['e e']:

(2) Essa é que é essa (e.g. Parente, 2005, p. 258).
(3) O meu é meu, o teu é nosso (e.g. Parente, 2005, p. 463).

Note-se que o número de variantes que um provérbio apresenta pode não estar relacionado com a sua disponibilidade lexical. Dito de outra forma, o facto de um provérbio apresentar muitas variantes não faz dele necessariamente um provérbio usual. Há provérbios muito usuais que apresentam uma tal fixidez dos seus constituintes que não admitem qualquer variação, como por exemplo o provérbio 'Contas são contas'. Em contrapartida, o provérbio 'Não se apanham trutas a bragas enxutas' tem pelo menos 12 variantes mas não é usual.

Desta forma, para podermos agrupar as diferentes variantes de provérbios numa mesma unidade paremiológica foi necessário utilizar outras metodologias, de que falaremos mais adiante neste trabalho.

Em complemento desta abordagem, efetuou-se uma seleção dos provérbios considerados mais usuais, percorrendo-se a base de dados de provérbios (Reis & Baptista, 2017a). Esta seleção foi feita por dois anotadores, os autores deste artigo, que são falantes nativos do português europeu, linguistas de profissão, e com vasto conhecimento de provérbios.

Tinha-se como hipótese subjacente a esta metodologia que os provérbios selecionados por ambos os anotadores seriam 'muito usuais' ('nível 2'), os selecionados por somente um dos anotadores seriam (apenas) 'usuais' ('nível 1') e os que não fossem assinalados por nenhum dos anotadores seriam 'não usuais' ('nível 0').

Verificámos uma elevada concordância entre os provérbios anotados, uma vez que 276 provérbios foram selecionados pelos dois anotadores - 'nível 2' ('Quem canta seus males espanta'); 566 provérbios foram selecionados por apenas um dos anotadores - 'nível 1' ('Em tempo de guerra não se limpam armas') e os restantes (+113.000), não foram selecionados por nenhum dos anotadores - 'nível 0'.

Contudo, estes números não são finais, uma vez que há repetições que terão de ser assinaladas e variantes de um mesmo provérbio que deverão ser agrupadas. Ainda assim, o facto de ter havido uma elevada concordância entre os anotadores poderá ser um indicador de que os provérbios escolhidos são efetivamente muito usuais.

De modo a podermos validar esta seleção, foi desenvolvido um inquérito por questionário, do qual falaremos de seguida.

O inquérito por questionário, a que chamámos 'Provérbios Usuais' [5] (Reis & Baptista, 2017a), teve como primeiro objetivo verificar se um conjunto de provérbios eram ou não conhecidos e utilizados pelos falantes nativos do português europeu. Para este inquérito foi selecionada uma amostra estratificada de 100 provérbios selecionados aleatoriamente, em três estratos. Desta amostra constam 25 provérbios de 'nível 2'; 25 de 'nível 1' e 50 de 'nível 0'. Era pedido aos informantes que indicassem para cada provérbio se conheciam e usavam ('nível 2'), se conheciam mas não usavam ('nível 1') ou se não conheciam ('nível 0').

Este questionário recebeu 738 respostas e permitiu-nos confirmar, em geral, a disponibilidade lexical que os anotadores atribuíram aos provérbios da base de dados.

Por um lado, pôde#se confirmar a seleção manual relativamente aos provérbios indicados como 'muito usuais' ('nível 2') e 'não usuais' ('nível 0'); por outro lado, os provérbios que tinham sido assinalados por apenas um dos anotadores - 'usuais' ('nível 1') - terão de ser analisados mais minuciosamente, de forma a se considerar se deverão integrar a lista dos provérbios 'muito usuais' ou passar para a classe dos 'não usuais'.

Com base nas respostas obtidas no questionário, calculou-se o índice de correlação de Pearson entre o nível indicado na resposta mais votada e o nível que lhe fora atribuído como referência (com base nos dois anotadores especialistas). Dito de outra forma, comparou-se a resposta mais votada pelos respondentes para cada provérbio com a classificação dos dois anotadores. Considerando os 100 provérbios do questionário, verificou-se que há uma correlação 'bastante alta' (0,77) entre o nível indicado na resposta mais votada e o nível que lhe fora atribuído como referência. Nesta interpretação do índice de correlação de Pearson, seguimos a proposta de Parsian (2015). Isto parece indicar que o nível atribuído pelos anotadores não difere substancialmente do que é atribuído pelos respondentes do inquérito, o que confirma, pelo menos parcialmente, a classificação inicial feita pelos anotadores para os restantes provérbios selecionados, sobretudo os de 'nível 1'.

Foram ainda utilizadas outras formas de validação da disponibilidade lexical dos provérbios que apresentámos neste questionário. É desses dados que falaremos em seguida.

Procurou-se o conjunto de provérbios apresentados no questionário de que falamos atrás nos motores de busca Google e Bing. A pesquisa restringiu-se a páginas do domínio de topo.pt e escritas em português europeu. Observou#se que os resultados, mais uma vez, confirmavam, no geral, a seleção manual dos anotadores, uma vez que os provérbios anotados como 'muito usuais' ('nível 2') eram efetivamente os que ocorriam com mais frequência na web. Pontualmente, verificou-se que alguns provérbios deste nível ocorrem um número reduzido de vezes, o que levou a reavaliar a sua classificação inicial (ou como de 'nível 1' e, mais raramente, como de 'nível 0'). Quanto aos provérbios de 'nível 0', considerados 'não usuais', confirma-se claramente que ocorrem muito poucas vezes ou não ocorrem de todo na web. A situação mais complexa verifica#se nos provérbios de 'nível 1', marcados como apenas 'usuais', já que alguns apresentam valores de frequência idênticos aos dos provérbios de 'nível 2' ('muito usuais') o que levou à análise mais cuidadosa das suas ocorrências e, por vezes, à sua reclassificação no 'nível 2'. Inversamente, certos provérbios de 'nível 1' parecem não ocorrer na web, o que poderá significar ser necessário avaliar melhor, por recurso a outros métodos, a sua disponibilidade lexical.

No geral, verifica#se que há uma correlação bastante alta entre os resultados da frequência dos dois motores de busca e a classificação inicial feita manualmente pelos anotadores.

Não sendo os motores de busca uma ferramenta suficientemente fiável, por si só, recorremos ao corpus de texto jornalístico CETEMPúblico (Santos & Rocha, 2001) para aferir a frequência destes provérbios. Foram utilizadas duas abordagens diferentes com base em ferramentas de processamento da linguagem natural, e

com recurso a transdutores de estados finitos (FST) (Paumier, 2016). Uma destas abordagens consistiu na construção manual de 'FST' que representam cada 'unidade paremiológica' (Figura 1).

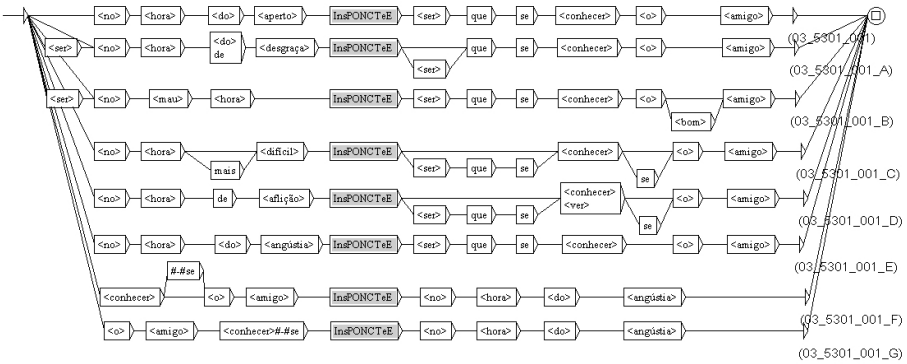

FIGURA 1.
Transdutor do provérbio "Na hora do aperto é que se conhecem os amigos" (e.g. Parente, 2005, p. 369).
Elaborada pelos autores.

Tentámos representar nestes FST todas as variantes de uma UP que encontrámos nas nossas fontes. Este é, no entanto, um processo em constante atualização, uma vez que poderão ser adicionadas outras variantes que não estejam atestadas na nossa base de dados. Esta abordagem tem a vantagem de poder recorrer à variação morfossintática, permitindo capturar assim todas as variantes que apresentem este tipo de variações ('Conhece-se o amigo' na hora da angústia; 'Conhecem-se os amigos' na hora da angústia). Em contrapartida, este método não permite capturar expressões que não estejam previstas no grafo. Tendo em conta que a construção destes transdutores é uma tarefa demorada e que ainda se encontra em curso, foi necessário empregar uma outra abordagem, mais expedita, que nos permitisse capturar os provérbios que ainda não tivessem sido abrangidos pela primeira. Esta consistiu no desenvolvimento de um conjunto de transdutores com base nas palavras-chave de cada provérbio (Figura 2). Este método tem como vantagem o facto de os 'FST' poderem ser gerados automaticamente a partir da base de dados. Ainda que não produza resultados tão precisos como a primeira abordagem, é um método com maior abrangência (recall).

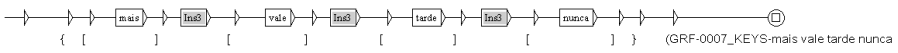

FIGURA 2.
Transdutor do provérbio "Mais vale tarde do que nunca" (e.g. Parente, 2005, p. 333).
Elaborada pelos autores.

O resultado da aplicação destes dois conjuntos de 'FST' no corpus CETEMPúblico (Santos & Rocha, 2001) permitiu-nos encontrar 30 provérbios diferentes, correspondendo a 13 unidades paremiológicas. Permitiu#nos ainda confirmar alguns dos provérbios mais utilizados e refletir sobre as variantes empregues. Todos os provérbios encontrados fazem parte do conjunto dos provérbios usuais ('nível 2'), de acordo com a referência da anotação manual.

Duas outras pesquisas foram realizadas tendo, de entre os objetivos estabelecidos, o de determinar os provérbios (e as suas variantes) mais difundidos em conjuntos de manuais escolares (Reis & Baptista, 2016b, 2017b) quer de português língua materna (LM), quer de português língua não materna (PLNM), com recurso a ferramentas de processamento de linguagem natural ('FST' que representam as unidades paremiológicas e 'FST' com as palavras-chave).

Duas expetativas orientaram as pesquisas: (i) por um lado, esperava-se que os provérbios usuais aparecessem repetidos ou no mesmo manual ou em diferentes manuais; (ii) por outro lado, julgava-se que

os provérbios presentes nos manuais corresponderiam, na grande maioria, a provérbios 'usuais' ou 'muito usuais', independentemente da sua frequência nestas obras.

A ideia por detrás desta segunda expetativa está relacionada com o facto de se pressupor que a escolha de provérbios realizada pelos autores destes manuais, de algum modo poderia refletir a sua disponibilidade lexical, já que é naturalmente desejável que os provérbios a apresentar neste tipo de instrumentos didáticos, dirigidos a um público de aprendentes, correspondam de forma mais direta a expressões quotidianas, a traços mais vincados da cultura portuguesa, ou seja, a provérbios usuais na língua e na cultura portuguesas.

Para a pesquisa de provérbios em manuais de português (LM), recorreu#se a um corpus de 38 manuais escolares, da Porto Editora, do 5 ao 12º anos, com 2,9 milhões de palavras, que já se encontrava digitalizado (Curto, 2014), tendo-se encontrado 272 provérbios, que correspondiam a 192 unidades paremiológicas distintas. Neste caso, não foi possível estabelecer uma lista dos provérbios mais usuais, uma vez que poucos se repetiram. Nos casos em que houve repetições, verificou#se que, por vezes, um mesmo provérbio era utilizado em mais do que um exercício no mesmo manual, o que poderá# ter influenciado ou enviesado os resultados obtidos.

Um outro aspeto a ter em consideração é o facto de alguns provérbios encontrados ocorrerem em textos literários que constam destes manuais ou serem referidos a propósito destes textos. Uma vez que os mesmos provérbios se poderão repetir em diferentes manuais escolares do mesmo ano de escolaridade, e dado que os textos estudados são os mesmos, essa repetição não é necessariamente significativa. Veja-se, por exemplo, o provérbio: "(4) Voz do povo, voz de Deus" (e.g. Garrido, Duarte, Rodrigues, Afonso, & Lemos, 2000, s.p.), que se encontra em quatro dos cinco manuais do 11º ano deste corpus, no texto dramático Frei Luís de Sousa, de Almeida Garrett "MARIA - Voz do povo, voz de Deus, minha senhora mãe: eles que andam tão crentes nisto, alguma cousa há-de ser (ato I, cena III)".

Ao compararmos os resultados de frequência dos provérbios nestes manuais com os níveis de disponibilidade lexical dos provérbios definidos manualmente pelos anotadores, verificámos que 50% dos provérbios encontrados são 'não usuais' (de 'nível 0'), 28% são 'muito usuais' (de 'nível 2') e 22% são apenas 'usuais' (de 'nível 1'). O número de provérbios 'não usuais' encontrado é igual à soma dos marcados como 'usuais' e 'muito usuais'. Este resultado não era o esperado, considerando que era desejável que os manuais representassem os provérbios (mais) usuais na língua e na cultura portuguesas. Ao verificarmos os índices de frequência destes provérbios 'não usuais' na web, pudemos confirmar que são efetivamente provérbios 'muito pouco usuais', uma vez que 26% não apresentaram quaisquer ocorrências e 66% destes provérbios tiveram menos de 30 ocorrências. Recorde#se que o limiar de frequência considerado para os provérbios de 'nível 2' na web era 50). Ainda assim, 8% dos provérbios encontrados apresentaram mais de 30 ocorrências, pelo que foram eventualmente reclassificados nos outros dois níveis de disponibilidade lexical ('nível 1', 'nível 2').

Um estudo semelhante foi desenvolvido num conjunto de manuais de português língua não materna (PLNM), publicados pela LIDEL, de vários níveis de proficiência, com aproximadamente de 250 mil palavras. Este conjunto de textos foi especificamente digitalizado para esta fase do estudo.

Os resultados desta pesquisa permitiram identificar 34 provérbios, correspondendo a 20 unidades paremiológicas distintas. À semelhança do que verificáramos nos manuais de português (LM), poucos provérbios se repetiram (apenas 7). Desta forma, não foi possível determinar se alguns provérbios seriam mais usuais nestes manuais. Pudemos, no entanto, observar que todos os provérbios encontrados podiam ser considerados 'usuais' ou 'muito usuais', com base nos dados de disponibilidade lexical, isto é, considerando a anotação manual que nos serviu de referência.

Outro questionário que desenvolvemos, o 'Questionário de avaliação da qualidade de jogos proverbiais' [6] (Reis, Pompili, Abad, & Baptista, 2017), consistiu na aplicação de 100 jogos – cloze questions - com provérbios (estímulos), que apresentavam quatro opções de resposta (uma resposta certa e três 'distratores'), todas linguisticamente motivadas. Destes jogos, 12 apresentavam duas respostas corretas. O questionário era constituído por duas partes distintas. Na 'Parte 1', apresentava-se o conjunto de jogos com provérbios que

se pedia que fossem resolvidos; na 'Parte 2', era solicitada a opinião dos respondentes relativamente à parte anterior.

O primeiro objetivo deste questionário consistia em avaliar a qualidade de um conjunto de jogos com provérbios, com vista a vir a inseri-los numa plataforma virtual de diagnóstico e terapia de algumas patologias da linguagem. No entanto, também nos permitiu validar, de modo indireto, a seleção dos provérbios usuais feita pelos dois anotadores e já descrita anteriormente, uma vez que a grande parte dos provérbios que serviram neste questionário (selecionados aleatoriamente e todos de 'nível 2') não integrara a amostra do questionário 'Provérbios Usuais' (apenas 21 se repetiram). Ao utilizarmos um outro conjunto de provérbios já assinalado como 'muito usuais', pudemos validar a classificação dos anotadores sempre que os jogos de provérbios foram corretamente resolvidos por um elevado número de respondentes.

Responderam a este questionário 164 pessoas e, em média, a taxa de acerto foi de 93%. Estes dados confirmam a apropriada classificação dos provérbios apresentados como 'muito usuais', pois em geral os falantes conheciam e sabiam usar os provérbios. Nos casos em que apresentámos mais do que uma solução correta, a maioria dos respondentes selecionou a opção que consideramos ser a variante mais usual dessa UP.

A 'Parte 2' do questionário permitiu#nos ainda validar a metodologia seguida na seleção dos 'distratores', tendo em vista uma eficiente construção dos exercícios. Em concreto, perguntava-se se a solução dos exercícios se encontrava entre as opções indicadas e se o informante conseguia completar facilmente o provérbio com base nas opções de resposta apresentadas, mesmo quando não o conhecia.

Na primeira questão, a alta percentagem de pessoas que indicou que a solução se encontrava 'sempre' ou 'a maioria das vezes' (53,7 e 44,5%, respetivamente) nas opções de resposta apresentadas, permitiu concluir, em geral, que os provérbios tinham sido reconhecidos pelos informantes. A segunda questão, quanto à facilidade em completar os provérbios com base nas opções de resposta apresentadas está ligada ao facto de essas respostas terem todas um racional subjacente, que permite determinar por analogia a resposta certa. Verificou#se que 20,7% dos respondentes selecionou a opção 'sempre' e 68,9% a opção 'a maioria das vezes', o que confirma a adequação dos 'distratores' produzidos.

Constituição da lista de provérbios usuais – Mínimo paremiológico

Os resultados obtidos em cada uma das etapas, já descritas anteriormente, permitiram-nos estabelecer a lista final de provérbios muito usuais - o mínimo paremiológico do português europeu -, que apresentamos em Anexo. De seguida, descreveremos com algum pormenor este processo de seleção.

Primeiramente, retomou-se a lista de provérbios que foi objeto de seleção manual pelos anotadores a partir da base de dados de provérbios. Foi necessário verificar todos os provérbios ou variantes repetidas que tinham sido selecionados inadvertidamente; eliminar as repetições; e escolher de entre as variantes repetidas a mais usual. Em alguns casos considerou#se apenas um dos membros dos provérbios, cuja frequência é muito superior à utilização da forma mais longa com os dois membros, como sucede, por exemplo, em "Não há regra sem exceção [nem mulher sem senão]" (e.g. Parente, 2005, p. 392). Parte deste trabalho já tinha sido feita, mas agora esta análise incidiu não só nos provérbios classificados como 'muito usuais' (de 'nível 2'), como também nos considerados 'usuais' (de 'nível 1').

Atentemos primeiro à lista de provérbios de 'nível 2'. Nesta fase, e sob um olhar mais crítico, resolvemos excluir desta lista 44 provérbios que tinham sido assinalados inicialmente pelos dois anotadores como muito frequentes. Esta decisão assentou em diferentes critérios: na baixa frequência de uso destes provérbios, que foi verificada na web, em corpora de texto jornalístico e de manuais escolares; e nos resultados obtidos pela aplicação de questionários, dos quais voltaremos a falar, de seguida.

A aplicação dos questionários permitiu-nos excluir, com segurança, alguns provérbios, como por exemplo: "(1) A preguiça morreu de sede à beira da água" (e.g. Parente, 2005, p. 49). Este provérbio, que consta do

questionário 'Provérbios Usuais' [7] (Reis & Baptista, 2017a), não foi reconhecido por 45,1% dos inquiridos, que selecionaram a opção 'Não conheço'.

Em alguns casos, a variante apresentada poderá ter tido alguma influência sobre as respostas fornecidas, apesar da advertência feita no início do questionário para que fossem consideradas outras variantes além da que se mostrava: 'interessa#nos apenas que nos indique se conhece e/ou usa os provérbios (ou uma das suas variantes) que lhe apresentamos'. Por essa razão, nos casos em que os resultados do questionário nos pareceram menos claros, complementámos esta informação de natureza introspetiva com dados de frequência na rede.

Nesta fase, a frequência de provérbios foi obtida com base no motor de busca Google, uma vez que já se tinha verificado que os resultados dos motores de busca Google e Bing eram idênticos/comparáveis. Uma vez mais, a pesquisa foi feita exclusivamente em páginas do domínio de topo.pt, e tendo em conta apenas as páginas escritas em português europeu (dados de janeiro de 2019). De modo a obtermos dados o mais fidedignos possível, percorremos até à última página os resultados de cada uma das pesquisas efetuadas pelo browser, o que nos possibilitou identificar os resultados mais relevantes, omitindo as entradas muito semelhantes às já apresentadas. Nestas pesquisas, tivemos em consideração quer a expressão ipsis verbis (digitando cada expressão entre aspas), quer as diferentes variantes desses mesmos provérbios (digitando cada uma das variantes manualmente ou substituindo certos elementos por wildcard), o que nos permitiu descartar a hipótese de o baixo uso de um dado provérbio poder dever-se à escolha da variante.

Por exemplo, considerando o provérbio "A preguiça morreu de sede à beira da água" (e.g. Parente, 2005, p. 49), somente encontrámos oito ocorrências do mesmo, ipsis verbis. Todas as variantes (conhecidas) deste provérbio têm uma fraca expressão na web, quase todas na variante acompanhada de 'andando a nadar', com a preposição 'a' ('morreu à sede') (#5 ocorrências), e uma com a preposição 'de' ('morreu de sede'). Encontrou-se ainda duas ocorrências de "A preguiça morreu de sede" (e.g. Machado, 1996, p. 50), sem um segundo hemistíquio. Não se encontrou quaisquer ocorrências de outras variantes já registadas.

Um caso particularmente expressivo é o provérbio: "(2) Não se apanham trutas a bragas enxutas" (e.g. Parente, 2005, p. 400) para o qual 91,5% dos respondentes indica 'Não conheço'. Este foi o único caso em que os informantes indicaram que não conheciam um provérbio, inicialmente marcado como sendo de 'nível 2'. Efetivamente, a frequência na rede do provérbio ipsis verbis é muito baixa (#3 ocorrências) e as variantes encontradas ocorrem predominantemente em listas de provérbios, enquanto outras variantes não ocorrem de todo.

Já o provérbio: "(3) Deus dá o frio conforme a roupa" (e.g. Machado, 1996, p. 190), apesar de ter sido marcado como 'muito usual', não era conhecido pela maior parte dos respondentes do questionário dos jogos, pois a maioria (68,9%) não completou corretamente o provérbio. Os 'distratores' apresentados ou exploravam o paralelismo formal e a rima: [deus dá o frio] 'e o diabo dá o brio' (45,7%); ou a rima: 'conforme o arrepio' (15,9%); ou a sinonímia: 'conforme a vestimenta' (7,3%). Apenas 31,1% acertou na opção correta.

Note-se que estes três provérbios aparecem em todas as coletâneas que utilizámos, apresentando inclusive, alguma variação. Face a estes resultados, transferimo-los para os provérbios de 'nível 1'.

Dos 44 provérbios retirados da listagem inicial de 276 de 'nível 2' (seleção manual), confirmámos que 35 apresentavam índices de frequência bastante baixos, mesmo quando se considerava o conjunto de variantes, por vezes numeroso, associadas à mesma 'unidade paremiológica'. Assim, por exemplo, para o provérbio: "(4) Cada um dança conforme a música"(e.g. Parente, 2005, p. 138) encontrámos 19 ocorrências na web, considerando todas as 9 variantes associadas a esta unidade paremiológica (a variante ilustrada em (4) ocorre 3 vezes), mas apenas uma variante parece mais usual: "Conforme se toca, assim se dança" (e.g. Parente, 2005, p. 172) (#13 ocorrências).

Por outro lado, para o provérbio: "(5) Antes a criança chore que a mãe suspire" (e.g. Parente, 2005, p. 85), que apresenta 2 variantes, apenas esta forma ocorre na rede e com uma frequência baixa (#6 ocorrências).

Como se pode verificar, o número de variantes de uma UP não está diretamente relacionado com a sua frequência em textos.

Removemos também da lista inicial dos provérbios de 'nível 2' cinco variantes das mesmas unidades paremiológicas, que se encontravam repetidas, veja-se, por exemplo:

(6) A descer todos os santos ajudam (e.g. Parente, 2005, p. 21).
(6a) Para baixo todos os santos ajudam (e.g. Parente, 2005, p. 512).

Retirámos também desta lista as expressões fixas:

(7) Os dados estão lançados (e.g. Parente, 2005, p. 496).
(8) Fia-te na virgem e não corras e verás o trambolhão que levas (e.g. Parente, 2005, p. 274).

Pois considerámos tratar-se de expressões, atualizáveis no discurso, e não propriamente provérbios. Removemos ainda: "(9) O bacalhau quer alho" (e.g. Parente, 2005, p. 439) por nos parecer sobretudo estar associado a uma canção popular e por não termos encontrado qualquer evidência do seu estatuto enquanto provérbio: todas as ocorrências que encontrámos ou estavam relacionadas com a canção popular ou só apareciam em listas de provérbios e não usados de forma integrada no discurso. Finalmente, retirámos a expressão: "(10) O rei vai nu" (e.g. Parente, 2005, p. 480) que, apesar da sua frequência bastante elevada, nos causou dúvidas relativamente à sua classificação como provérbio. A grande maioria das ocorrências por nós encontradas de "O rei vai nu" (e.g. Parente, 2005, p. 480). correspondem a referências explícitas ao conto de Hans Christian Andersen (2006), de onde esta expressão deriva. No entanto, encontrámos alguns exemplos de emprego da expressão sem qualquer relação explícita ao conto, num uso que se parece bastante com o de muitos provérbios: "É que apesar [de os] donos da cultura em Portugal se estarem a divertir bué, 'o rei vai 'mesmo' nu'. [8]

Tal como foi necessário ajustar a lista de provérbios de 'nível 2', também foi indispensável efetuar alterações na lista de provérbios de 'nível 1'. Assim, 76 provérbios (dos 566) que tinham sido selecionados por apenas um dos anotadores foram reclassificados, tendo agora sido integrados no mínimo paremiológico, como, por exemplo, "Da discussão nasce a luz" (e.g. Parente, 2005, p. 182).

Algumas das expressões incluídas no MP poderão surpreender o leitor por serem slogans e/ou citações, alguns até com autor conhecido, como, "Há mar e mar, há ir e voltar" (slogan cunhado por Alexandre O'Neill para a campanha de sensibilização do Instituto de Socorros a Náufragos, para prevenir os afogamentos nas praias portuguesas) ou "Tudo está bem quando acaba bem" (tradução do título da peça All's well that ends well de Shakespeare - (Shakespeare, 2009, s.p). Consideramos que estas expressões já são parte do fundo comum de conhecimento da língua, e que, em relação a elas, a noção de autoria já terá praticamente desaparecido, daí que sejam muitas vezes utilizadas proverbialmente. Vejamos, a propósito, os seguintes exemplos retirados da web:

> O ditado popular que afirma que '[...] há mar e mar, há ir e voltar [...]', é volta e meia 'violado' pelos pescadores minhotos, que acabam por morrer na faina, como hoje aconteceu em Vila Praia de Âncora, Caminha [9] .
>
> Chegamos ao fim de mais um ano com altos e baixos, mais positivo para uns do que outros, mas 'como diz o ditado', 'tudo está bem quando acaba bem', e por isso é importante acabar 2016 em festa, como é costume, celebrando a passagem de ano.

Os exemplos aqui apresentados estão associados a algumas das fórmulas que muitas vezes se usam para introduzir um provérbio no discurso: 'como diz o ditado'. O facto de, por um lado, estas expressões, de que até conhecemos a autoria, serem consideradas e usadas como provérbios pelos falantes; e, por outro lado, a sua elevada frequência de uso, tal como se reflete no número de ocorrências que apresentam na web; levaram-nos a incluí-las no MP. Note-se, todavia, que, até por eventual desconhecimento nosso, poderá haver nesta mesma lista outras expressões que são, de facto, citações mas que não foram por nós identificadas como tal.

Ainda em relação a este conjunto de 76 provérbios, verificámos que alguns deles também ocorriam nos manuais tanto de língua materna, como de língua não materna, por exemplo, "Quem conta um conto

acrescenta um ponto" (e.g. Ferreira, & Bayan, 2016, s.p.). (#7 ocorrências) e "Uma imagem vale mais que mil palavras" (Pinto, Baptista & Fonseca, 2004, s.p.). (#4 ocorrências). Tal pode ser considerado como uma confirmação da seleção no caso dos manuais de língua não materna. Já nos manuais de língua materna, esta evidência não é tão significativa.

Verificámos que a maioria dos provérbios de 'nível 1' que foram reclassificados para 'nível 2' surge em todas as coletâneas da nossa base de dados. Há, no entanto, algumas exceções que gostaríamos de referir. O provérbio "Nada dura para sempre" (Parente, 2005, p. 371). ocorre apenas em uma das coletâneas. Ao verificámos a sua frequência na web, constatámos que ocorre 98 vezes (sem resultados repetidos), o que consideramos uma alta frequência de uso quando comparada com a frequência de muitos outros provérbios usuais. Confirmámos ainda que esta expressão é reiteradamente usada de modo proverbial e identificada explicitamente como um provérbio, como sucede em:

> 'Nada dura para sempre....' Foi esta 'a máxima que repeti' muitas vezes, foi esta 'a máxima que passei a outras pessoas'. 'Nada dura para sempre', o amor magoa, a vida é lixada e tantas outras coisas negras e cheias de sofrimento [11].

Os dados obtidos através da aplicação dos questionários também contribuíram para esta reclassificação, na medida em que muitos destes provérbios foram reconhecidos por um elevado número de respondentes. No caso do provérbio "Na hora do aperto é que se conhecem os amigos" (e.g. Parente, 2005, p. 369), que surge no questionário 'Provérbios Usuais' [12] (Reis & Baptista, 2017a), 94,6% dos respondentes conhece o provérbio. Este provérbio apresenta diversas variantes na base de dados, surgindo em todas as coletâneas deste estudo. Daí a sua reclassificação no 'nível 2'.

O provérbio "Quem tem unhas é que toca guitarra" (e.g. Machado, 1996, p. 555). também foi reclassificado tendo em conta os resultados dos questionários (este é um dos casos de provérbios que se repetiram em ambos os questionários). No questionário 'Provérbios Usuais' (Reis & Baptista, 2017a) 78,4% dos inquiridos conhece este provérbio. No caso do questionário dos jogos, 81,1% selecionou a opção correta.

Após todas as reclassificações efetuadas e ao refletirmos sobre a lista de provérbios assim estabelecida, verificámos que havia quatro outros provérbios com uma frequência bastante elevada e que deveriam ter sido incluídos no Mínimo Paremiológico do Português Europeu. São estes os provérbios: "Ano novo, vida nova"; "Contra factos não há argumentos"; "O que é bom acaba depressa" e "O ataque é a melhor defesa', cujo número de ocorrências na web é 175, 91, 107 e 72, respetivamente. Constatámos que, talvez por lapso, estes não haviam sido selecionados por nenhum dos anotadores.

Considerações finais

Em conclusão, ainda que qualquer seleção tenha sempre uma componente de arbitrariedade, os cuidados metodológicos de que nos rodeámos, bem como a justificação explícita da forma como eles foram aplicados, levam-nos a considerar que este estudo produziu uma lista de provérbios muito conhecidos e muito usuais que pode ser utilizada como o mínimo paremiológico do português europeu (MP). O estudo introduz também o conceito de unidade paremiológica (UP), formalizando explicitamente a relação entre provérbios e suas variantes e delimitando o conceito de variação.

Este instrumento poderá ter diversas aplicações, por exemplo para o desenvolvimento de instrumentos de diagnóstico ou terapia de certas patologias da linguagem; ou para a aprendizagem de português, tanto como língua materna como língua estrangeira.

Recentemente em (Reis, Baptista, & Mamede, 2020), repetiu-se a experiência relativa ao cálculo da frequência de provérbios e respetivas variantes no corpus jornalístico CETEMPúblico (Santos & Rocha, 2001), usando-se a cadeia de processamento da linguagem natural (PLN) STRING - Statistical and Rule-Based Natural Language Processing Chain (Mamede, Baptista, Diniz, & Cabarrão, 2012) [13]. Foi

assim possível validar a seleção de parte dos 318 provérbios que constam do mínimo paremiológico aqui estabelecido. Trata#se de um tipo de texto que, como refere Figueiredo (2012), nos dias de hoje não é particularmente propenso a apresentar provérbios. Ainda assim, encontrámos 457 padrões diferentes que correspondem a 909 ocorrências de provérbios. Do conjunto das unidades paremiológicas registadas no MP, 115 não foram encontradas neste corpus. Confirmou#se ainda a seleção das variantes de provérbios selecionadas para identificar as UP no MP, uma vez que as variantes que aparecem neste corpus correspondem também às variantes mais usuais.

Utilizamos também as mesmas gramáticas locais da STRING para processar as bases de dados de provérbios Mosaico (Cercifat, 2002) [14] e do Projeto Natura (1994) [15], disponíveis na rede. Realce-se que, intencionalmente, estas coletâneas não foram usadas para constituir a nossa base de dados de provérbios para que pudessem agora servir de corpus de teste. Os resultados confirmam que os provérbios selecionados para o MP são realmente expressões usuais porque, em grande parte, surgem em ambas as coletâneas. Na primeira base de dados (Mosaico), que tem 5.363 entradas, foram encontrados 358 provérbios, correspondendo a 269 unidades paremiológicas diferentes. Do MP, 49 provérbios não foram encontrados nesta base de dados ('A carne é fraca'). Na segunda base de dados (Natura), que continha originalmente 2.293 entradas correspondentes a provérbios, foram desdobradas as entradas que tinham variantes e retiradas as repetições, resultando em 1.902 entradas diferentes. A STRING permitiu-nos encontrar aqui 407 provérbios, que correspondem a 236 UP diferentes. Há, no entanto, 82 provérbios no MP que não constam da lista de provérbios do Projeto Natura (1994) ("A curiosidade matou o gato"). Das unidades paremiológicas encontradas nestas duas bases de dados, 221 provérbios são comuns a ambas. Os resultados obtidos parecem, pois, validar a seleção manual de provérbios muito usuais.

Num trabalho futuro, pretendemos completar as gramáticas locais da STRING, acrescentando às expressões regulares já construídas as variantes que eventualmente ainda não tenham sido representadas. Pretendemos ainda continuar a atualizar este mínimo paremiológico, com recurso a blogues interativos e a miniquestionários a serem aplicados na web.

Agradecimentos

Esta investigação foi parcialmente suportada por fundos nacionais através da Fundação para a Ciência e a Tecnologia (ref. UIDB/50021/2020).

Material suplementar

Anexo (png)

Referências

Andersen, H. C. (2006). *The complete Hans Christian Andersen fairy tales*. New York: Gramercy Books.

Cercifat. (2002). Mosaico. *Base de dados de provérbios*. Recuperado de http://www.cercifaf.org.pt/ mosaico.edu/ 1c/ proverb1.txt

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

Costa, J. (1999). *O livro dos provérbios portugueses*. Lisboa, PT: Editorial Presença.

Curto, P. (2014). *Classificador de textos para o ensino de português como segunda língua* (Dissertação de Mestrado). Instituto Superior Técnico, Universidade de Lisboa, Lisboa.

Ďurčo, P. (2004). Slovak proverbial minimum: the empirical evidence. In C. Földes (Ed.), *Res humanae proverbiorum et sententiarum. Ad honorem Wolfgangi Mieder* (p. 59-69). Tübingen, DE: Narr.

Ďurčo, P. (2005a). Empirisch- und Korpusbasierte Untersuchungen der Sprichwörter. *Zeitschrift für germanistische Sprach# und Literaturwissenschaft in der Slowakei, 1*(3), 47-57.

Ďurčo, P. (2005b). Paremiologické minimum slovenčiny. Vy#sledky a porovnania. In R. Blatná, & V. Petkevič (Eds.), *Jazyky a jazykov # eda. Sborni#k k 65. narozeninám prof. PhDr. Františka Čermáka, DrSc* (p. 45-61). Praha, CZ: U#stav Českého národni#ho korpusu.

Ďurčo, P. (2006). Methoden der sprichwortanalysen oder auf dem weg zum sprichwörter#optimum. In A. H. Buhofer, & H. Burger (Eds.)*, Phraseology in motion I. Methoden und kritik* (Phraseologie und Parömiologie, n. 19, p. 3-20). Baltmannsweiler, DE: Schneider Hohengehren.

Ďurčo, P. (2007). Paremiologické optimum slovenčiny. In W. Chlebda (Ed.)*, Frazeologia a j #ezykowe obrazy s#wiata przełomu wieków* (p. 171-177). Opole, DE: Uniwersytet Opolski.

Ďurčo, P. (2015). Empirical research and paremiological minimum. In H. Hrisztova-Gotthardt, & M. Aleksa Varga (Eds.), *Introduction to paremiology: a comprehensive guide to proverb studies* (p. 183-205). Berlin, DE: De Gruyter.

Ďurčo, P., & Meterc, M. (2014). Empirične paremiološke raziskave tipov ekvivalentnosti in suprasemantičnih razlik v slovenščini in slovaščini. *Slavia centralis, 6*(2), 20-36.

Ferreira, A., & Bayan, H. (2016). *Na Onda do Português 3*. Lisboa, PT: Lidel.

Figueiredo, G. (2012). *O gênero proverbial na imprensa: usos e funções retóricas* (Tese de Doutorado). Universidade Federal de Pernambuco, Recife.

Garrido, A., Duarte, C., Rodrigues, F., Afonso. F., & Lemos, L. (2000). *Práticas 11*. Lisboa, PT: Fundação Calouste Gulbenkian.

Gasanova, M. A., & Taibova, L. Y. (2016). Revisiting paremiological units of the Tabasaran language. *Oriental Studies, 25*(3), 99-105. Doi: 10.22162/2075-7794-2016-25-3-99-105

Machado, J. P. (1996). *O grande livro dos provérbios*. Lisboa, PT: Editorial Notícias.

Mamede, N., Baptista, J., Diniz, C., & Cabarrão, V. (2012). STRING: an hybrid statistical and rule-based natural language processing chain for Portuguese. In *Processing of the 10th International Conference on Computational Processing of the Portuguese Language* . Coimbra, PT. Recuperado de https://www.inesc-id.pt/ficheiros/publicacoes/8578.pdf

Meterc, M. (2012). Online questionnaire providing information on most well#known and well#understood proverbs in slovene language. In *Paper presented at the EUROPHRAS Conference*. Maringor, SI. Doi: 10.13140/RG.2.1.1491.7203

Meterc, M. (2014). *Primerjava paremiologije v slovenskem in slovaškem jeziku na osnovi paremiološkega optimuma* (doktorska disertacija). Filozofska fakulteta Univerze v Ljubljani. Ljubljana.

Mieder, W. (1992). Paremiological minimum and cultural literacy. In S. J. Bronner (Ed.). *Creativity and tradition in folklore: new directions* (p. 185-203). Logan, UT: Utah State University Press.

Mieder, W. (1994). *Wise words (RLE folklore)*. London, UK: Routledge.

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

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

Parsian, M. (2015). *Data algorithms: recipes for scaling up with hadoop and spark*. Sebastocol, CA: O'Reilly Media.

Paumier, S. (2016). *Unitex 3.1 - user manual*. Université de Paris- Est,Marne-la-Vallée, FR: Institut Gaspard Monge.

Permjakov, G. L. (1973). On the paremiological level and paremiological minimum of language. *Proverbium, 1*(22), 862-863.

Pinto, E., Baptista, V., & Fonseca, P. (2004). *Plural 11*. Lisboa, PT: Porto Editora.

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* (p. 427-446). Universidade do Algarve, Faro.

Reis, S., & Baptista, J. (2016b). O uso de provérbios no ensino de português. In Processing of the *10th Interdisciplinary Colloquium on Proverbs* (p. 521-538). Tavira, PT.

Reis, S., & Baptista, J. (2016c). Portuguese proverbs: types and variants. In G. C. Pastor (Ed.), *Computerised and corpus-based approaches to phraseology: monolingual and multilingual perspectives* (p. 208-217). Geneva, CH: Editions Tradulex.

Reis, S., & Baptista, J. (2017a). Estimating lexical availability of European Portuguese proverbs. In R. Mitkov (Ed.), *Computational and corpus#based phraseology. EUROPHRAS 2017. Lecture Notes in Computer Science, 10596* (p. 232-244). Londres, UK: Springer International Publishing.

Reis, S., & Baptista, J. (2017b). Os provérbios em manuais de ensino de português língua não materna. In V. Pinheiro, & G. H. Paetzold (Eds.), *STIL 2017. XI Brazilian Symposium in Information and Human Language Technology and Collocated Events. Proceedings of the Conference* (p. 247-255). Uberlândia, MG.

Reis, S., Baptista, J., & Mamede, N. (2020). Processing Proverb Variation in (European) Portuguese: Integrating the Paremiological Miminum in STRING. In *Paper presented in 2nd Workshop on Linguistic Tools and Resources for Paraphrasing in Portuguese, collocated with International Conference on the Computational Processing of Portuguese (PROPOR 2020)*. Colégio do Espírito Santo, Évora, PT. Doi: 10.13140/RG.2.2.14935.62880

Reis, S., Pompili, A., Abad, A., & Baptista, J. (2017). O provérbio como estímulo num terapeuta virtual. In *6º Simpósio Mundial de Estudos sobre o Português (Simpósio 77 - A Importância da aprendizagem lexical*. Lisboa, PT.

Ruiz-Ayu#car, M., & Sevilla Muñoz, J. (2016). *El mi#nimo paremiológico: aspectos teóricos y metodológicos* (Biblioteca fraseológica y paremiológica, Série mi#nimo paremiológico, n. 1). Madrid, ES: Centro Virtual Cervantes, Instituto Cervantes.

Santos, D., & Rocha, P. (2001). Evaluating CETEMPúblico, a free resource for portuguese. In*Proceedings of the 39th Annual Meeting of the Association for Computational Linguistic*s (p. 442-449). Toulouse, FR. Recuperado de http:www.linguateca.pt/CETEMPublico/

Shakespeare, W., (2009). *All's Well that Ends Well: The Cambridge Dover Wilson Shakespeare* (S. Quiller-Couch & J. Dover Wilson, Eds.). Cambridge, UK: Cambridge University Press. doi:10.1017/CBO9780511704079

Anexo

Mínimo Paremiológico do Português Europeu

Notas

[1] Informação recuperada de http://www.ileel.ufu.br/ileel/?p=15619 33

[2] Recuperado em https://cvc.cervantes.es/lengua/refranero/

[3] Esta opção não está contemplada no nosso estudo, tendo apenas sido apresentadas as opções de resposta 'Conheço e uso', 'Conheço mas não uso', 'Não conheço' (logo, não uso). Esta opção (iii) introduz uma variável que, quanto a nós, seria difícil controlar: a representação do significado do provérbio.

[4] Recuperado em www.Korpus.sk.

[5] Recuperado de https://goo.gl/forms/NvfkLXMTdCb5dA0t2

[6] Recuperado de https://goo.gl/forms/H0NXEKTxUBxFFZzB2

[7] Disponivel em https://goo.gl/forms/NvfkLXMTdCb5dA0t2

[8] https://www.jornaltornado.pt/quando#o#rei#vai#nu#e#preciso#tomar#a#palavra/

[9] https://tvi24.iol.pt/naufragio/vila-praia-de-ancora/ha-mar-e-mar-ha-ir-e-nao-voltar

[10] http://lolesmtg.pt/tag/passagem#de#ano/

[11] Recuperado de https://wrestling.pt/beyond-the-mat-27-last-forever/

[12] Recuperado de https://goo.gl/forms/NvfkLXMTdCb5dA0t2

[13] Recuperado de http://string.l2f.inesc-id.pt

[14] Recuperado de http://www.cercifaf.org.pt/mosaico.edu/1c/proverb1.txt

[15] Recuperado de https://natura.di.uminho.pt/~jj/pln/proverbio.dic