Provérbios portugueses usuais: distribuição em corpora

Sónia Reis[1], **Jorge Baptista**[1,2], **Nuno Mamede**[1,3]

[1]HLT – Human Language Technologies Lab
Lisboa, Portugal

[2]Universidade do Algarve, Faculdade de Ciências Humanas e Sociais
Faro, Portugal

[3]INESC ID Lisboa
Lisboa, Portugal

reis.soniamm@gmail.com, jbaptis@ualg.pt, nuno.mamede@inesc-id.pt

Abstract. *Proverbs are a special type of linguistic unit that have been largely ignored by the Natural Language Processing (NLP) community, though they pose interesting challenges to NLP systems. This paper presents the procedure of integrating the Paremiological Minimum of Portuguese into the STRING system, and the distribution of these most usual proverbs in three different (European) Portuguese corpora.*

Resumo. *Os provérbios são um tipo especial de unidades linguísticas que tem sido amplamente ignorado pela comunidade de Processamento de Linguagem Natural (PLN), apesar de levantarem desafios interessantes para o processamento. Este artigo apresenta o procedimento de integração do Mínimo Paremiológico do Português no sistema STRING e a distribuição desses provérbios mais usuais em três corpora distintos do português (europeu).*

1. Processamento de provérbios em textos

Tanto quanto sabemos, os provérbios estão praticamente ausentes dos recursos linguísticos de muitos sistemas de Processamento de Linguagem Natural (PNL) desenvolvidos para o português. Tal resulta, provavelmente, do seu estatuto linguístico pouco consensual, algures entre o léxico e a cultura, que está certamente na origem da sua ausência dos dicionários de língua, encontrando-se antes recenseados em coletâneas de provérbios, onde convivem com outras expressões (locuções) de natureza muito variada, frequentemente idiomáticas (i.e. semanticamente não composicionais). Outro aspeto é o seu extenso número e a sua variação formal, lexical e sintática, que dificultam o reconhecimento destas expressões em textos [Rassi et al. 2014a, Rassi et al. 2014b]. Finalmente, o seu estatuto de *citação*, nem sempre formalmente assinalado (aspas/discurso relatado), confere-lhe uma certa autonomia relativamente o texto em que se insere, permitindo que sejam violados mecanismos de coesão discursiva (processos anafóricos, concordância verbal na subordinação), ainda que os seus elementos se prestem a ser retomados noutros locais do discurso, dando origem a criativos jogos de palavras. São, pois, complexos os desafios que os provérbios levantam ao seu processamento automático em textos.

O estudo dos provérbios no quadro do PLN pode dar origem a aplicações societalmente relevantes. [Reis and Baptista 2016b] desenvolveram dois conjuntos de jogos

com provérbios, linguisticamente motivados, com base nos provérbios do MP, que podem ser utilizados para ensino de língua ou até para o dignóstico/terapia de algumas patologias da linguagem. Mais recentemente, [Mendes and Oliveira 2020a] testaram diferentes técnicas de representação semântica (e.g. Jaccard) para avaliar a similaridade semântica entre um *corpus* de aproximadamente 1.600 provérbios e manchetes de jornais, no quadro de uma tarefa de recomendação automática de texto. A avaliação dos resultados por meio de questionário revelou que, na maioria das vezes, as pessoas conseguiram estabelecer uma relação entre a expressão selecionada e a manchete correspondente, chegando, mesmo, a achá-la potencialmente engraçada. Verificam ainda que os provérbios que partilham as mesmas palavras com a manchete, nomeadamente os escolhidos por métodos mais simples (e.g. Jaccard) são mais facilmente relacionados com as manchetes. Já o uso de representações semânticas mais profundas como *word embeddings* (e.g. BERT) [Mendes and Oliveira 2020b] revelou piores resultados, o que foi justificado pelos autores pela linguagem figurada própria destas expressões. Numa linha mais próxima da deste artigo, [Davis et al. 2021] analisam a frequência e a dinâmica de provérbios em diferentes tipos de texto e ao longo do tempo, baseando-se numa lista de +14 mil provérbios americanos e usando como *corpora*: (i) o *corpus* Gutenberg (60.000 documentos); (ii) o *corpus* New York Times (1,8 milhões de artigos de 1987-2007); (iii) dados do Google (a partir de 2020); e (iv) dados do Twitter (a partir de janeiro de 2021). Os autores identificaram as expressões que mais se repetem em cada um destes *corpora*, admitindo que uma limitação do seu estudo é a questão da representatividade, já que os dados de partida, de um dicionário de provérbios americano, são limitados. Ainda assim, a maior disponibilidade de dados textuais abre caminho a estudos fraseológicos longitudinais.

Enquanto unidade linguística, dada a sua variação, os provérbios podem ser organizados em *unidades paremiológicas* (UP), isto é, unidades conceptuais representadas pelo conjunto das múltiplas variantes de um mesmo provérbio e que são definidas com base em critérios formais, semânticos e pragmáticos [Reis 2020]. Dado o elevado número de provérbios conhecidos, na ordem das dezenas de milhar, pode ser útil restringir o âmbito da investigação ao *Mínimo Paremiológico* (MP) de uma língua, ou seja, o conjunto dos provérbios mais usuais de uma dada língua, conhecidos e empregues pela maioria dos falantes de uma dada comunidade linguística.

Até à data, os provérbios têm estado fora do âmbito dos objetos linguísticos processados pela STRING [Mamede et al. 2012], uma cadeia de Processamento de Linguagem Natural, construída especificamente para processar textos em português. Este artigo apresenta a integração na STRING do conjunto de expressões que formam o *Mínimo Paremiológico do Português Europeu* [Reis and Baptista 2020], e está organizado do seguinte modo: na secção seguinte, apresentamos o mínimo paremiológico e a forma como foi constituído; depois, descrevemos o método de representação dos provérbios do MP e suas variantes para a sua integração na STRING; de seguida, descrevemos os 3 *corpora* utilizados neste estudo e a distribuição dos provérbios do MP nesses textos, procurando determinar fatores que expliquem as assimetrias encontradas; o artigo conclui-se com breves notas sobre perspetivas futuras.

2. Mínimo paremiológico

O *Mínimo Paremiológico do Português Europeu* (MP) [Reis and Baptista 2020] foi determinado, reunindo primeiro várias listas de provérbios a partir de dicionários/coletâneas

de referência [Moreira 1996, Costa 1999, Parente 2005, Machado 2011] e depois estimando a sua *disponibilidade lexical* com base na sua ocorrência em variadas fontes [Reis and Baptista 2016a]. Em primeiro lugar, com base numa listagem de +114 mil provérbios foi produzida, de forma independente, uma classificação manual por dois anotadores humanos, que identificaram os provérbios que consideravam serem "usuais", sendo depois comparadas as respetivas anotações. Os resultados obtidos permitiram atribuir um nível provisório de disponibilidade lexical aos provérbios, utilizando uma escala de 3 níveis (0, 1 e 2, em que '0' corresponde a expressões pouco usuais, raramente disponíveis; '1', moderadamente disponíveis; e '2' para as expressões muito usuais, altamente disponíveis). De seguida, a partir de uma seleção aleatória de provérbios de cada um dos níveis assim determinados, foi aplicado um questionário *on-line*, a que responderam 735 informantes, o qual veio confirmar, em grande medida, a seleção dos anotadores. Por outro lado, foi determinada a frequência dessa mesma lista de provérbios obtida a partir da *web*, no domínio de topo `.pt` e utilizando dois motores de busca (Bing e Google). Os resultados de ambos os motores de busca foram muito semelhantes (Pearson: 0.96), e mostram uma correlação relativamente alta (Pearson: aprox. 0.70) com a classificação manual dos anotadores. A mesma experiência foi realizada no *corpus* CETEM*Público* [Santos and Rocha 2001] e os resultados obtidos correlacionaram-se bem com a classificação manual dos anotadores, já que não há provérbios de nível 0 (raros ou pouco usuais) e que foram encontrados mais casos de provérbios de nível 2 (muito usuais) do que de nível 1 (usuais). Uma vez que os provérbios são frequentemente utilizados num contexto educacional para a aprendizagem de língua (e cultura), foi reproduzida a experiência tanto em manuais didáticos de português L1 [Reis and Baptista 2016c] como em manuais de português L2 [Reis and Baptista 2017]. Os resultados destas experiências tornaram possível o estabelecimento do *Mínimo Paremiológico do Português Europeu*, que contém 318 UPs, ou seja os provérbios mais usuais e representando mais de 14.500 expressões proverbiais (variantes) [Reis and Baptista 2020]. Por outras palavras, cada provérbio está associado ao conjunto das suas variantes, já recolhidas em dicionários e coletâneas especializadas e, em alguns casos, variantes encontradas em outras fontes, incluindo a Internet.

3. Expressões regulares e integração do MP na STRING

Descrevemos agora a metodologia de integração do léxico do Mínimo Paremiológico no sistema de processamento computacional do português STRING [Mamede et al. 2012], com vista à identificação destes provérbios nos textos em que ocorram. Para tal, determinaram-se primeiro os requisitos de informação necessários para encontrar a solução ótima com vista à integração deste tipo de objetos na arquitetura geral da STRING. Um dos requisitos fundamentais é poder representar os provérbios contando com a análise morfossintática dos itens lexicais, no mínimo, a informação quanto ao lema das formas e a respetiva categoria morfossintática. Outro requisito prende-se com a possibilidade de inserção de separadores no meio das expressões, já que a pontuação de muitas expressões proverbiais é bastante "criativa". Assumiu-se, além disso, que, à semelhança de uma palavra composta, os provérbios funcionam como uma "ilha" no corpo do texto, não sendo relevante (pelo menos numa análise inicial) atribuir-lhes uma análise sintática, proceder à extração de papéis semânticos [Talhadas et al. 2013], ou mesmo fazer a resolução de relações anafóricas [Mitkov 2002, Marques 2013]. Finalmente, ainda que nesta fase apenas se tenha representado os provérbios do Mínimo

Paremiológico, constituído pelas 318 unidades paremiológicas mais usuais e frequentes, pretende-se que o método permita, progressiva e cumulativamente, representar uma mais lata extensão do rico património linguístico e cultural que constitui o léxico dos provérbios (dezenas de milhar de unidades paremiológicas).

Assim, considerando a arquitetura da STRING, decidiu-se delimitar e anotar o provérbio numa fase bastante inicial do processamento, tratando-o como uma unidade textual, mas só depois da fase de análise e desambiguação morfossintática, para poder recorrer a essa informação. Isso corresponde a utilizar o mecanismo que, no analisador sintático da STRING, se chama de *gramáticas locais* (*GramLoc*). Estas consistem em, numa fase preliminar de análise sintática (ing. *parsing*), construir um nó `noun` na estrutura da frase, com base no emparelhamento de uma dada expressão com um padrão descrito por uma *expressão regular* (*RegExp*), associando-lhe depois as propriedades linguísticas relevantes. No sistema, as expressões regulares das gramáticas locais obedecem a uma sintaxe própria, cujo formalismo é razoavelmente complexo, pelo que a sua construção manual, diretamente a partir das formas a descrever, está muito sujeita à introdução de erros. Por essa razão, a integração dos provérbios na STRING foi feita em duas etapas: num primeiro momento, as variantes dos provérbios são representadas por meio de expressões regulares relativamente transparentes, cuja construção é mais simples e mais adequada para um linguista, preocupado sobretudo em descrever os padrões combinatórios dos provérbios, do que o formalismo usado pelo sistema nas suas gramáticas locais; e num segundo momento, um programa especialmente construído para o efeito converte essas expressões regulares no formalismo mais complexo das gramáticas locais usado pela STRING.

Dada a natureza da variação formal observada neste tipo de expressões, adotou-se uma representação por expressões regulares que permitisse descrever a estrutura e variação destas formas tendo em conta: (i) possibilidade de indicação dos items lexicais tanto pela *forma* como pelo <*lema*>, e.g. `gato` ou `<gato>`, eventualmente acompanhada da categoria morfossintática, e.g. `<ser,V>`; (ii) possibilidade de representar qualquer sequência de separadores (‘`#`’) ou de palavras (‘`@`’); (iii) operadores mais usuais de expressões regulares como a disjunção ‘`|`’ ou a sequência vazia ‘`$`’. Assim, por exemplo, considerando o provérbio *Velhos são os trapos* (MP-315), em que se observa a possibilidade de comutar *trapos* com *farrapos*, o linguista constrói a RegExp:

```
velhos <ser,V> os (trapos|farrapos).
```

que é depois convertida automaticamente no formalismo da GramLoc:

```
1> noun[proverb=+]
@= ?[surface:velhos];?[surface:Velhos],
verb[lemma:ser], ?[surface:os],
?[surface:trapos];?[surface:farrapos].
```

Esta GramLoc constrói um nó `noun`, com a propriedade `proverb=+`, juntando a sequência de palavras formadas pelas formas *velhos*, uma forma flexionada associada ao lema do verbo *ser*, o artigo *os* e as formas em alternativa (`;`) *trapos* ou *farrapos*; note-se que a primeira forma tem de ser representada como duas alternativas, para dar conta do emprego de maiúscula inicial (uma alternativa seria usar o lema `<velho>`). Num segundo momento, ao nível da análise sintática do texto, uma regra geral extrai a

dependência unária PROVERB, que identifica o provérbio:

```
| noun#1[proverb] | PROVERB(#1).
```

O sistema identifica, assim, este provérbio nos textos em que ocorra, produzindo como saída a dependência correspondente (exemplo retirado do *corpus* Desportivo, v. adiante; na dependência os elementos do provérbio são, para já, identificados pelo respetivo lema):

> **[DSP]** *Cumpriu o seu 300º jogo para o campeonato e continuou a provar que* velhos são os trapos.
> `PROVERB(velho ser o trapo)`

Para cada unidade paremiológica do MP foram construídas manualmente mais de 1.100 RegExp, dando conta de fenómenos tão variados como permutas, redução de elementos, variação lexical e inserção de sinais de pontuação. Em paralelo, foram produzidos semiautomaticamente mais de 14.500 exemplos ilustrativos das variantes representadas por essas RegExp, que funcionam como material de validação das GramLoc que são construídas a partir daquelas.

4. Distribuição dos provérbios do MP em *corpora*

O estudo da distribuição dos provérbios do *Mínimo Paremiológico do Português Europeu* (MP) em *corpora* visa não só avaliar o desempenho da STRING em textos de natureza e tópicos variados, como também procurar padrões que revelem diferenças no uso destas expressões em contextos diferentes. Para aferir a distribuição em *corpora* dos provérbios usuais do MP, foram utilizados 3 *corpora* já existentes e processados pela STRING: o *CETEMPúblico* (CTP) [Santos and Rocha 2001], o *Desportivo* (DSP) e o *Parlamento* (PRL) [Trindade 2020]. O *corpus* CETEMPúblico, de cariz jornalístico, é um *corpus*, disponível publicamente, que contém 175.350.145 palavras (após processamento na STRING) e é distribuído pela Linguateca[1]. O *corpus Parlamento* resulta da compilação das atas de sessões da Assembleia da República, de 1976 a 2018, apresentando 123.633.859 palavras. Finalmente, o *corpus Desportivo*, com 100.161.374 palavras, é um *corpus* de texto jornalístico, de temática desportiva (sobretudo futebol), composto por textos dos jornais *O Jogo* (1999-2005) e *A Bola* (2000-2006).

Através da aplicação das gramáticas locais da STRING aos 3 *corpora*, foram identificadas 7.334 ocorrências de provérbios, das quais, 45,1% encontram-se no Desportivo, 30,4% no CETEMPúblico e 24,5% no Parlamento. Por razões de espaço, apenas alguns resultados serão reportados neste artigo[2].

A grande maioria das expressões encontradas nos *corpora* correspondem efetivamente a instâncias dos provérbios do MP, sendo a precisão alcançada bastante elevada (99,86%). A avaliação da abrangência, pelo menos para já, só faria sentido para os provérbios do MP, já formalizados, e respetivas variantes. Um pequeno número de casos (#10) são, porém, *falsos-positivos* e correspondem às seguintes situações: (i) *inserções*, como sucede na RegExp: `($|a) paciência @ <ter,V> limites.` que descreve o

[1] `www.linguateca.pt/cetempublico`

[2] A lista integral das UP, com a distribuição dos provérbios por cada *corpus*, bem como outras informações quanto às respetivas variantes, encontra-se disponível *on-line*: https://www.researchgate.net/publication/354997837; DOI: 10.13140/RG.2.2.14732.44164.

provérbio **MP_024** *A paciência tem limites*, pelo que o sistema emparelha esta expressão com a frase seguinte:

[CTP] [. . .] *José Afonso mostra que a festa de_os sons não* **tem limites** .

(ii) expressões ambíguas como, por exemplo:

[CTP] *Só que o problema não é* **querer , é poder**.
[CTP] *Deve situar se acima de os partidos políticos , sem* **querer ser contrapoder** *nem em relação a o Governo , nem em relação a a Assembleia da República*
[CTP] *E , sugere , o ‘ salve se* **quem puder ’ poderá** *não ser assim tão mau .*
[DSP] *E a deslocação a o American Airlines Arena não será o melhor jogo para* **quem procura encontrar** *definitivamente o caminho de as vitórias .*

No primeiro exemplo, estamos perante sequências de palavras de duas orações distintas, separadas por vírgula, que emparelharam com a RegExp do provérbio **MP_290** *Querer é poder*, pois nela admitia-se a inserção de pontuação (‘#’): `querer# <ser,V> <poder,V>.` No segundo exemplo, mais curioso, o nome *contrapoder*, cujo lema na STRING é *poder*, também foi capturado pela expressão regular. No exemplo seguinte, a expressão *Salve-se quem puder* está integrada como discurso reportado (entre aspas) e como sujeito de *poderá*, pelo que emparelha com a RegExp: `quem <poder,V># <poder,V>.` do provérbio **MP_271** *Quem pode, pode*. No segundo caso, estamos perante uma construção de verbo auxiliar modal *procurar* + infinitivo que se confunde com esta variante do provérbio **MP_272** *Quem procura acha*.

Em termos de diversidade de UP distintas presentes em cada *corpus*, observa-se valores semelhantes: o CTP apresenta ocorrências de 241 UP diferentes, o DSP 230 e o PRL apenas 200. Considerando como *densidade* (*Dens*) o n.º de palavras de cada *corpus* a dividir pelo nº de provérbios nele encontrados, verifica-se que Dens(CTP)=78.562, Dens(DSP)=30.279 e Dens(PRL)=68.915, ou seja, o *corpus* CTP apresenta uma maior densidade destas expressões (2,6 vezes relativamente ao DSP e 1,1 vezes mais em relação ao PRL). A densidade média dos provérbios do MP no conjunto dos 3 *corpora* é de 1 provérbio por cada 54.424 palavras. Quando comparamos as séries de valores de frequência das unidades paremiológicas encontradas em cada *corpus*, verifica-se uma correlação média-alta entre os dados obtidos nos *corpora* CTP e DSP (Pearson=0,691) e entre o CTP e o PRL (Pearson=0,671) mas uma correlação média-baixa entre o DSP e o PRL (Pearson=0,379). Neste sentido, a distribuição dos provérbios nos *corpora* DSP e PRL é mais dissemelhante do que a que se verifica quando cada *corpus* é comparado com o o CTP. Finalmente, verifica-se que metade (159; 50%) dos provérbios do MP ocorrem nos 3 *corpora*; cerca de um quarto (77; 24%) ocorre apenas em 2 dos *corpora* (a maioria no CTP e no DSP: 48; 15%); 40 provérbios (13%) ocorrem apenas num dos *corpora*, distribuindo-se de forma equilibrada; e apenas 42 não ocorreram em nenhum dos 3 *corpora*. A Fig. 1 apresenta (por ordem de ID) a distribuição dos 10 provérbios mais frequentes do MP pelos 3 *corpora* e que representam 25% do total de provérbios encontrados. Ora, apesar de, neste conjunto dos provérbios mais frequentes, a maioria (50%) das ocorrências se encontrar no DSP, verifica-se claras assimetrias da distribuição de cada provérbio.

Estes provérbios parecem também ser mais neutros do ponto de vista dos tópicos a que se podem aplicar e respetivos contextos de uso em que podem ocorrer (exceto, talvez, o **MP_185** *O segredo é a alma do negócio*). Veja-se, por exemplo, alguns dos contextos em que ocorre o provérbio **MP_133** *Mais vale tarde do que nunca*, que é, quanto ao total de ocorrências, o mais frequente nestes 3 *corpora*:

[CTP] *Depois de um complexo processo burocrático , a inauguração de uma embarcação para o transporte de 22 passageiros em o passado mês de Agosto acabaria por ser saudada com um*

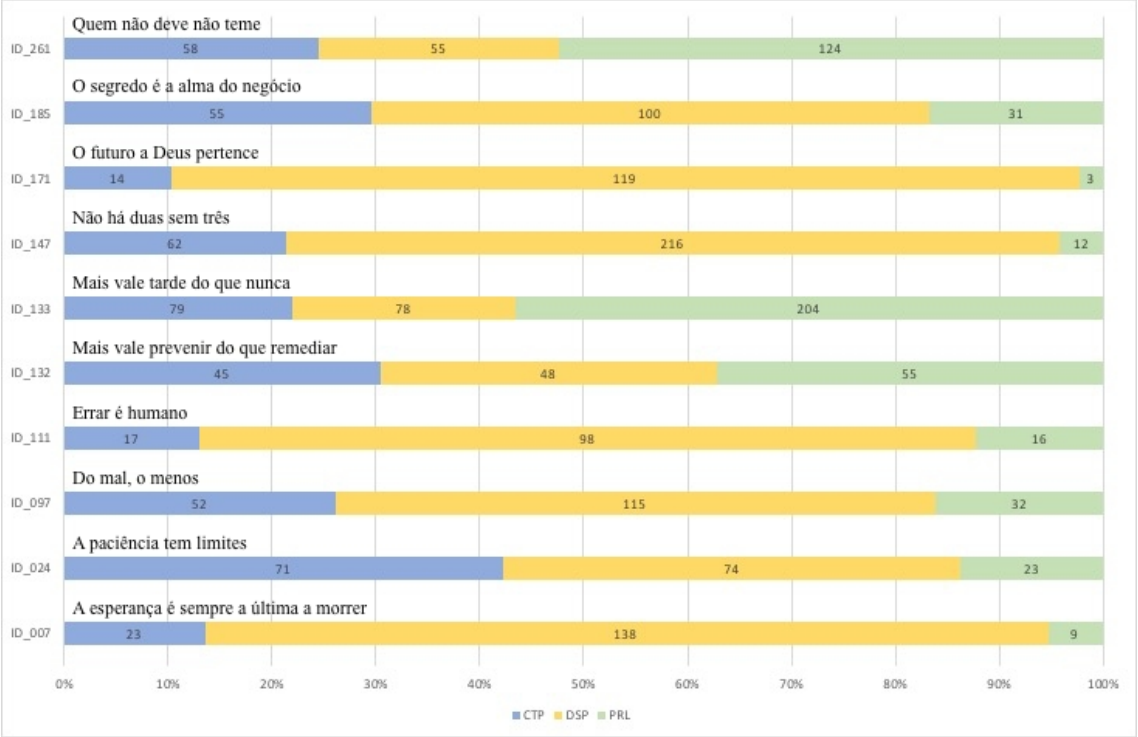

Figura 1. Distribuição dos 10 provérbios mais frequentes do MP pelos 3 *corpora*.

"**mais vale tarde do que nunca**".
[DSP] Antes tarde que nunca *: em a segunda parte , Scolari colocou em campo os dois Juninhos (Pernambucano e Paulista) e com isso ganhou maior eficácia em o meio-campo .*
[PRL] *Sr. Presidente , Sr. Primeiro-Ministro , em primeiro lugar , relativamente a as duas medidas que anunciou , era importante lembrar que* "**mais vale tarde do que nunca**" .

A aparente neutralidade temática deste provérbio deveria permitir que o mesmo ocorresse em diferentes situações comunicativas. Contudo, relativamente à sua distribuição nos 3 *corpora*, verifica-se que este provérbio ocorre praticamente o mesmo número de vezes no CETEMPúblico e no Desportivo (79 e 78 vezes), mas aparece 204 vezes no Parlamento. Considerando *frequência média* (*Frqmed*), como o n.º ocorrências do provérbio dividido pelo n.º total de ocorrências de todos os provérbios num dado *corpus*, temos para este provérbio: Frqmed(CPT)=3,53%, (DSP)=2,35% e (PRL)=11,37%, ou seja, o provérbio é 3 a 5 vezes mais frequente no PRL do que nos outros dois *corpora*, no CTP e no DSP, respetivamente. Como explicar esta assimetria? Veja-se, também, o que se passa com os provérbios com maior assimetria neste conjunto: **MP_171** *O futuro a Deus pertence*, em que 119 (88%) das ocorrências se encontram no *corpus* DSP, e **MP_007** *A esperança é sempre a última a morrer* com 170 (81%) instâncias no mesmo *corpus* DSP. Ambos os provérbios têm uma frequência baixa no *corpus* CTP (14 e 23 ocorrências) e praticamente residual no PRL (3 e 9 ocorrências), respetivamente. Dificilmente é possível detetar alguma razão no conteúdo e nos contextos de uso destes provérbios que justifique esta distribuição.

Dos 318 provérbios do MP, há 42 (13,2%) que não aparecem em nenhum dos 3 *corpora*. Trata-se, nalguns casos, de provérbios tematicamente bastante mais marcados do que os da lista acima, o que poderá explicar a sua ausência neste tipo de textos: um ideal de mulher: **MP_020** *A mulher e a sardinha querem-se da mais pequenina*; a alimentação: **MP_016** *A laranja, de manhã é ouro, à tarde é prata e à noite mata*; os filhos: **MP_114** *Filhos criados, trabalhos dobrados*; etc. Noutros casos, são provérbios que ditam uma filosofia de vida, de ordem geral, e.g. **MP_065**

Cada qual com o seu igual; **MP_159** *No meio é que está a virtude*; **MP_204** *Os opostos atraem-se*; **MP_247** *Quem espera sempre alcança*; **MP_275** *Quem sai aos seus não degenera*; **MP_316** *Vivendo e aprendendo*, sendo mais difícil explicar por que razão não ocorreram numa quantidade tão apreciável de textos. A maioria (237, 74%) dos provérbios do MP aparece com uma frequência inferior a 50 ocorrências. Alguns destes são também tematicamente bastante marcados, como, por exemplo, **MP_011** *A fome é o melhor tempero*, **MP_035** *Abril, águas mil*, ou **MP_117** *Gordura é formosura*. Outros, nem por isso: **MP_032** *A vida são dois dias*, **MP_042** *Amigo não empata amigo*, **MP_088** *Deitar cedo e cedo erguer, dá saúde e faz crescer*. Destes, 24 provérbios ocorrem apenas uma vez: **MP_148** *Não há regra sem exceção*, **MP_175** *O primeiro milho é dos pardais*, **MP_177** *O que arde cura*.

5. Conclusão

Neste artigo, apresentámos sucintamente os desafios que pela sua natureza, variação formal e modo de funcionamento nos textos, os provérbios levantam ao seu processamento computacional. De seguida, delineámos a metodologia seguida na constituição do Mínimo Paremiológico do Português Europeu (MP), constituído pelos 318 provérbios mais usuais, de acordo com critérios de frequência em *corpora* de natureza variada e de disponibilidade lexical. Os provérbios estão organizados em Unidades Paremiológicas (UP), que agregam a um mesmo provérbio o conjunto de variantes que este permite, recolhidas de diversas fontes. Em seguida, descrevemos o modo como este léxico de provérbios e suas variantes foram formalizados em gramáticas locais para integrar a STRING, de modo a permitir a sua identificação em textos. Finalmente, usando os resultados do processamento, descrevemos a distribuição dos provérbios em 3 *corpora* de natureza distinta e de grande dimensão.

A aplicação das gramáticas locais aos 3 *corpora* permitiu a identificação com elevada precisão (99,8%) de mais de 7.300 expressões proverbiais, nas suas múltiplas variantes. O método apresentado é, pois, válido e pode ser aplicado a textos de diversa natureza. Confirma-se, talvez dada a natureza destes *corpora*, a baixa frequência dos provérbios em textos escritos. Os poucos casos de falsos-positivos encontrados (#10) resultam sobretudo ou de inserções de elementos lexicais, ou de sinais de pontuação espúrios, ou ainda de problemas gerais de ambiguidade. O estudo da distribuição dos provérbios pelos três *corpora* revela a diversidade de situações e contextos de usos, dada a diversidade de unidades paremiológicas encontradas em cada *corpus*. Contudo, não é possível discernir padrões consistentes que expliquem, dada a natureza dos textos, as assimetrias encontradas na distribuição de cada unidade paremiológica pelos 3 *corpora* utilizados, seja quanto à sua frequência média, seja quanto às temáticas a que se aplicam. Este resultado é, em parte, esperável, já que, tratando-se dos provérbios mais usuais, seria esperável que apresentassem uma distribuição mais lata. Este artigo permitiu, assim, a construção de um conjunto de textos com provérbios usuais anotados, que ficará disponível para a comunidade científica. No futuro, a exploração de outro tipo de fontes textuais, ainda que suscite certas cautelas metodológicas na sua validação, deverá permitir encontrar outros contextos e novas variantes, validando eventualmente a intuição de que se trata de um tipo de unidade linguística mais ligada a um registo informal e a contextos predominantemente orais. Outra dimensão a explorar é a dinâmica temporal dos provérbios, sobretudo quando é possível relacionar estes com os tópicos (ou mesmo os acontecimentos) que motivaram o seu emprego, como sucede nos *corpora* do Desportivo e do Parlamento. Finalmente, pretende-se integrar a informação sobre os provérbios nos *corpora* já processados, permitindo a pesquisa *on-line* no demonstrador da STRING [Trindade 2020][3].

Agradecimentos. Parte da investigação para este artigo foi financiada por fundos públicos, pela Fundação para a Ciência e a Tecnologia (UIDB/50021/2020).

[3] `http://string.hlt.inesc-id.pt`

Referências

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

Davis, E., Danforth, C. M., Mieder, W., and Dodds, P. S. (2021). Computational paremiology: Charting the temporal, ecological dynamics of proverb use in books, news articles, and tweets. http://arxiv.org/abs/2107.04929.

Machado, J. (2011). *O Grande Livro dos Provérbios*. Casa das Letras, (4ª ed.), Alfragide.

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

Marques, J. (2013). Anaphora resolution. Master's thesis, Instituto Superior Técnico - Universidade de Lisboa, L^2F/INESC-ID, Lisboa.

Mendes, R. and Oliveira, H. G. (2020a). Comparing different methods for assigning Portuguese proverbs to news headlines. In Mikolov, T., Yih, W.-T., and Zweig, G., editors, *Linguistic regularities in continuous space word representations. Proceedings of NAACL-HLT, NAACL.*, pages 746–751. 11th International Conference on Computational Creativity (ICCC'20), ACL.

Mendes, R. and Oliveira, H. H. (2020b). TeCo: Exploring Word Embeddings for Text Adaptation to a given Context. In *Proceedings of ICCC*. 11th International Conference on Computational Creativity (ICCC'20), ACL.

Mitkov, R. (2002). *Anaphora Resolution*. Pearson – Prentice Hall.

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

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

Rassi, A. P., Baptista, J., and Vale, O. A. (2014a). Proverb variation: Experiments on automatic detection in Brazilian Portuguese texts. In Baptista, J., Mamede, N., Candeias, S., Paraboni, I., Pardo, T., and Volpe Nunes, M., editors, *Computational Processing of the Portuguese Language*, volume 8775 of *Lecture Notes in Computer Science / Lecture Notes in Artificial Intelligence*, pages 141–152, Berlin. 11th International Conference PROPOR'2014, São Carlos – SP, Brazil, October 8-10, 2014, Springer.

Rassi, A. P., Vale, O. A., and Baptista., J. (2014b). Automatic detection of proverbs and their variants. In Pereira, M., Leal, J., and Simões, A., editors, *Proceedings of the Symposium on Languages, Applications and Technologies (SLATE'14)*, pages 235–250, Leibniz (Germany). Symposium on Languages, Applications and Technologies (SLATE'14), Bragança (Portugal), June 19-20, 2014., Schloss Dagstuhl - Leibniz-Zentrum fur Informatik, Dagstuhl Publishing.

Reis, S. (2020). *Expressões proverbiais do português: Usos, variação formal e Identificação automática.* PhD thesis, Universidade do Algarve, Faro, Algarve, Portugal.

Reis, S. and Baptista, J. (2016a). Estimating lexical availability of european portuguese proverbs. In Mitkov, R. and Corpas Pastor, G., editors, *EUROPHRAS 2017*, volume 10596 of *Lecture Notes in Computer Science*, pages 232–244, Cham. Springer.

Reis, S. and Baptista, J. (2016b). Let's Play with Proverbs? NLP Tools and Resources for iCALL Applications around Proverbs for PFL. In *Proceedings of the International Interdisciplinary Conference in Social and Human Sciences*, Faro, Portugal. University of Algarve, Faculty of Economics.

Reis, S. and Baptista, J. (2016c). O uso de provérbios no ensino de português. In Soares, R. & Lauhakangas, O. (Eds.) *10th Interdisciplinary Colloquium on Proverbs*, Actas ICP16 Proceedings. Tavira: AIP-IAP, 2017, pp. 521–538.

Reis, S. and Baptista, J. (2017). Os provérbios em manuais de ensino de português língua não materna. In Vládia Pinheiro & Gustavo Henrique Paetzold (Eds.) *Proceedings of Symposium in Information and Human Language Technology* Uberlandia, MG, Brazil, October 2-5, 2017, Sociedade Brasileira de Computação, pp. 247–255.

Reis, S. and Baptista, J. (2020). Determinação de um mínimo paremiológico do português europeu. *Acta Scientiarum. Language and Culture*, 42(2):e52114. https://doi.org/10.4025/actascilangcult.v42i2.52114.

Santos, D. and Rocha, P. (2001). Evaluating CETEMPúblico: A Free Resource for Portuguese. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 442–449, Toulouse, France.

Talhadas, R., Baptista, J., and Mamede, N. (2013). Semantic roles annotation guidelines. Technical report, L2F/INESC ID Lisboa.

Trindade, J. (2020). Syntax Deep Explorer: Integrating multi-corpora support into a corpus analysis tool. Master's thesis, Instituto Superior Técnico, Universidade Técnica de Lisboa, L2F/INESC-ID, Lisboa.