# Estimating Lexical Availability of European Portuguese Proverbs

Sónia Reis[1] and Jorge Baptista[1-2]

[1] University of Algarve, Campus de Gambelas 8005-139 Faro, Portugal
[2] L2F/INESC-ID, Lisbon, Portugal
reis.soniamm@gmail.com jbaptis@ualg.pt

**Abstract.** This paper relates data on lexical availability with data on textual frequency of proverbs in European Portuguese. Each data source should provide different perspectives on the use of proverbs in the language. This should allow an empirically well-motivated selection of proverbs aiming at the development of NLP resources, specifically for applications for learning Portuguese as a Foreign Language and for the diagnosis/therapy of speech impairments/disabilities. A large database (over 114,000 proverbs and their variants) was independently classified by two annotators, according to intuitively estimated lexical availability. Next, a random, stratified sample was selected and lexical availability was then confirmed with an online survey. Frequency data was gathered from two web browsers and a large-sized, publicly available, *corpus* of journalistic texts. Results from the survey, the web and the *corpus* by and large confirm the initial intuitive classification and a core of commonly used proverbs was defined.

**Keywords:** European Portuguese Proverbs, Frequency in *corpus*, Lexical availability.

## 1 Introduction: Using Proverbs

Proverbs are used frequently and in many communicative contexts [1]. In spite of their colloquial/popular status and (mostly) oral transmission process [2], they are found not only in oral communication but also, though perhaps less frequently, in written discourses, and in different text types and genres [3], serving several types of rhetorical functions within discourse [4]. Because of their rich cultural and linguistic content [5], proverbs have been used in many applications, namely, as linguistic material for language learning [6] and language impairment diagnosis or speech therapy [7].

For language learning, proverbs provide a wide spectrum of expressive effects and a cross-cultural perspective on the language community and their symbolic heritage [2]. Furthermore, they are concise, often highly figurative, linguistic structures, naturally yielding to syntactic as well as culturally-oriented exploration in pedagogic context [8,9].

As linguistic material for diagnosis/therapy of language impairment, particularly in the case of pathologies resulting from trauma, proverbs are deemed as effective tools,

since they can be used as stimuli or prompts to exercise different cognitive structures, particularly long-term memory, even when the ability to speak is impaired [10–14].

Therefore, proverbs have been used in speech therapy, for example in tasks requiring the patient to complete a proverb, to explain its meaning, or the conditions in which it would be adequate to use it.

In spite of the crucial role that the adequate use of proverbs plays in the development of such exercises or tasks, little is said about their selection. Particularly in the case of didactic games involving proverbs, many available exercises could perfectly have used other linguistic material and no proper justification of the selected proverbs is provided [15].

Concerning proverbs, the selection of adequate material by specialist from these two areas faces several difficulties [16]. Selected proverbs should be commonly known, in order to reflect in a representative way, the culture of the language community using them. On the other hand, vocabulary involved in those expressions cannot be too rare or unknown. Furthermore, the choice of phraseological material should provide relevant items for the pedagogic goals of the exercises/activities [15].

As stimuli for speech therapy or language impairment diagnosis, selection of adequate examples is crucial. It is not easy to determine whether the fact of the patient not recognising the proverb, or the failure in producing an adequate answer to a fill-in-the-blank task, is due to a pathologic condition (e.g. dementia), or it is just due to the fact that he/she does not recognise/know the proverb.

Concerning Portuguese, and more specifically European Portuguese, [17] report several on-line resources, already available for learning Portuguese as a Foreign Language (PFL). The *Ciberescola da Língua Portuguesa*[1] and the *Centro Virtual Camões*[2] make available a set of didactic games, some of them involving proverbs. These games consist, basically, in completing a proverb or explaining its meaning.

For speech therapy/diagnosis, on the other hand, there are very few, publicly available, virtual therapists. One of them is the VITHEA system (Virtual Therapist for Aphasia Treatment) [18], which aims at the treatment of aphasia, featuring different types of visual and auditory stimuli, especially for eliciting vocabulary. To date, only a small number of exercises with proverbs have been produced.

*Lexical availability* is the key concept concerning these selection requirements. Besides other issues that may be task- or domain-specific (that should be used either in language learning or in speech therapy), a lexically available set of proverbs, consists of expressions:

a) that are easily recognised as such by the linguistic community as whole;
b) whose meaning and pragmatic conditions of use are widely known by native speakers;
c) that occur in a broad set of communicative situations.

[1] http://www.ciberescola.com/, last accessed 2017/05/13.
[2] http://cvc.institutocamoes.pt/, last accessed 2017/05/13.

Linguistic items presenting such requirements are considered to be lexically available, in the sense that they are part of the shared knowledge of the linguistic community, in as much the same way as the meaning and syntax of a commonly used verb is known, with high likelihood, by any native speaker of that given language. Lexically available items are also deemed to show a significant frequency on the language daily use, so that frequency can be viewed as an indirect signal of that availability.

The problem at hand is, thus, a question of devising the appropriate method for selecting a representative sample from the *corpus* of proverbs available for a given language. This problem is somewhat similar to the definition of a common vocabulary from the large lexicon of a language [19]. However, because of their primary oral mode of transmission and their colloquial nature, finding evidence of proverbs' use in written *corpora* is not a trivial task: in several languages, such as Portuguese, many large-sized available *corpora* are built from journalistic texts, and writing and style conventions strongly advise against using such colloquial expressions [20][3]. Besides, proverbs often present lexical and syntactical *variation* [21], which renders the task of finding them in texts much more complex than just looking for simple lexical items (words or phrases).

In view of the above, this paper aims at establishing, on solid empirical grounds, a core set of very commonly used proverbs, widely recognised (and adequately interpreted) by the majority of the (European) Portuguese, native speaking community. This selection could then be used in different scenarios, such as in language teaching and speech therapy, among other practical applications.

The remainder of the paper is structured as follows: Section 2 presents the method for manually selecting an initial set of proverbial expressions; then, using a stratified sample, classified according to their estimated lexical availability, that selection was validated using a survey. Next, in Section 3 the frequency of that sample is obtained from two popular web browsers, in order to further validate the initial selection. Finally, in Section 4 those same proverbs were queried in a large-sized, publicly available, European Portuguese *corpus*. Frequency data from the two types of sources (web and *corpus*) are compared against the estimated lexical availability, in order to produce an empirically well-motivated selection of commonly used proverbs that may be reliably used for different applications. The paper concludes (Section 5) by presenting the main findings and suggestions for future work.

## 2 Lexical Availability and Proverb Selection

In order to tackle the selection of an initial set of proverbs, candidate to the status of lexically available items, a large data base with 114,413 proverbs and their variants was used [22]. These were collected from four dictionaries of European Portuguese, which were digitised and then manually corrected. Each proverb was given a unique identifier (ID), indicating its source. After removing the stop words, each proverb was associated to a set of keywords: full verbs, nouns and adjectives, for the most part.

[3] static.publico.pt/nos/livro estilo/13-rigor-e.html, last accessed 2017/05/13.

Two annotators both native speakers of European Portuguese and extensive knowledge of the proverbial stock of the language independently marked the proverbs they recognised and deemed as usual. Annotator 1 marked 739 proverbs, while Annotator 2 marked 379 proverbs. This produced a tiered list of proverbs, ranked by levels (0 to 2). In total, 276 expressions were considered usual by both annotators (level 2), 566 proverbs were considered usual by only one of the annotators (level 1), and the remainder forms (113,571) have not been marked by either (level 0).

The initial assumption is that level 2 proverbs are lexically highly available expressions, level 1 are less so (only moderately available), and level 0 items, constituting the bulk of the database, though they are part of the *corpus* of proverbs of the language, are not sufficiently usual (seldom available) to be included in a selection aiming at the applications envisaged in this paper (language learning and speech therapy, for example).

Since only two annotators were involved, it was then deemed necessary to further confirm their selection. Furthermore, since a significant mismatch was found between the two annotations, it should be ascertained which proverbs from level 1 (where the two annotators did not agree) should be integrated in level 2, or left as only moderately available (level 1), or even removed from the selection altogether (and integrated in level 0).

A survey was thus built to confirm this initial selection. However, because the total set of proverbs (and variants) from levels 1 and 2 is too large to be presented in a survey to a wide audience, for it would require a long time to be answered by each participant; a random, stratified sample of the list of proverbs was produced, namely 50 items from levels 1 and 2 (25 from each), and 50 items from level 0. Because of the random selection, the items were manually revised in order to avoid repetition of proverbs from two different sources or using two variants of the same proverb.

Google Forms was used to build the survey and collect the answers. Some personal data was collected to characterise the sample: gender (M/F/undisclosed), age (less than 18, 18-30, 31-50, over 50), school level (basic, secondary, university level, other), nationality (short answer) county of residence (Portugal's 22 counties, including the autonomous regions of Madeira and Azores), area of residence (urban/rural). The detailed analysis of this data is presented elsewhere [16]. The selected proverbs were presented in random order, but always in the same order to every participant. For each proverb, the participant was asked to indicate whether he/she did not know proverb, or knew it but did not use it, or knew it and used it. These answers correspond to the 0 to 2 levels of the tiered selection of the random sample.

The survey was divulged among the authors' list of contacts, both individuals and groups, potentially reaching over 3,600 people. The survey was open for 7 days, and 735 answers were gathered before closing the survey (answer rate: 20,4%).

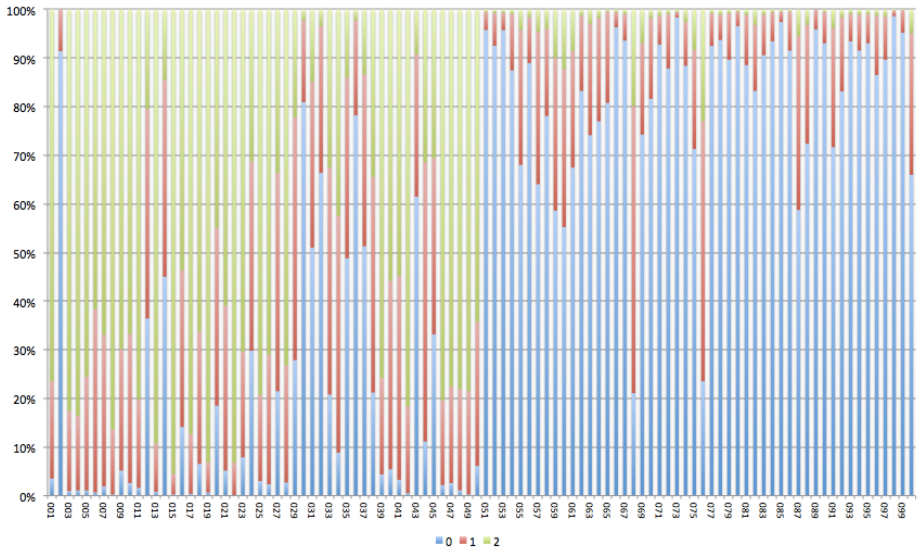


**Fig. 1.** Fig. 1. Assigning lexical availability to proverbs. Horizontal axis: the sample of 100 proverbs, identified by their ID code; from ID-001 to ID-025: reference level 2 (highly available); from ID-026 to ID-050: reference level 1 (moderately available); from ID- 051 to ID-100: reference level 0 (seldom available). Vertical axis: Percentage of answers from the survey (N=735) as they map onto the reference values: 2: I know and use this proverb; 1: I know but I do not use this proverb; 0: I do not know this proverb.

Figure 1 shows the results from the survey. It can be seen that most answers for the level 2 proverbs (ID-001 to ID-025) were recognised and marked by the subjects, either as in the same level 2 (known and used) or in level 1 (known but not used). A small number of cases do not follow this pattern: In the case of proverb:

ID-002 *Não se apanham trutas a bragas enxutas*
'Trouts cannot be cached/fished with dry trousers'

most participants did not know the proverb and only 8% knew it but did not use it. This proverb includes the archaic word *bragas* 'trousers', which may be one of the reasons for it not being recognized. In fact, common variants of the proverb replace this disused noun either for the modern equivalent *calças* 'trousers' or by a phonetically similar noun *barbas* 'beards', which does not change the overall figurative meaning of the expression. For another four cases, the sum of answers indicating levels 1 and 2 is above 50%:

ID-012 *São mais as vozes que as nozes*
'There are more voices then nuts',

ID-014 *A preguiça morreu de sede à beira da água*
'The sloth died of thirst (sitting) by/next to the water',

ID-020 *A morte não escolhe idades*
'Death chooses no ages', and

ID-024 *Hoje por mim, amanhã por ti*
'Today for me, tomorrow for you'.

In some cases, this can also be a result of the random selection of these proverbs/variants. In the case of proverb ID-020, a much more common variant exists with the noun *amor* 'love', which was confirmed by the queries on the web: the frequency ratio *morte*/*amor* 'death/love' in Google is 40/70 (0.57), and in Bing 40/237 (0.17).

On the other hand, among the many variants of proverb ID-024, there is one where 1st- and 2nd-person pronouns switch places:

*Hoje por ti, amanhã por mim*
'Today for you, tomorrow for me'.

This variant (an instance of the so-called 'golden rule') is also almost as frequent as the one shown in the survey. As the survey's variant constitutes an 'inversion' of the golden rule, this may be the cause for the lower availability level assigned by the survey. To sum up, selecting the most lexically available variants of a proverb proves to be almost as much important as choosing the proverbs themselves, in view of defining their lexically availability.

The situation is somewhat fuzzier in the case of level-1 proverbs (ID-026 to ID-050). Notice that this group of proverbs corresponds to a disagreement between the two annotators, as only one selected them as lexically available. Again, the (random) choice of a variant can be the cause for the proverb not being recognised. In the case of proverb:

ID-043 *Morra o gato, morra far*to
'May the cat die, may it die fully satisfied'

the variant

*Morra Marta, morra farta*,

with the proper noun *Marta* 'Martha', is much more frequent. Queries on the web yielded a *gato*/*Marta* ratio of 3/38 (0.08) in Google and 2/59 (0.03) in Bing. The remaining cases caution for a careful review of the full list of level-1 proverbs (and their variants), using frequency data (from the web and eventually other sources) to support this classification.

Finally, all level-0 proverbs (ID-051 to ID-100) were also assigned the same lexical availability level by the subjects, except for two proverbs (ID-068 and ID-076), since a level 1 was assigned instead.

This data seems to confirm in general the manual assignment of 0- and 2-level of lexical availability to the sample of proverbs. The cases from level 1 must be considered with care, as results do not show a clear-cut distinction between this level and the other two.

## 3 Proverbs in the web

The frequency of the same sample of proverbs was obtained from querying two web browsers, Google and Bing. The query was restricted to the exact matches, within the Portugal top domain (.pt), and selecting only pages written in European Portuguese. The matches were manually perused for false positives. Fig. 2 shows these results. It is clear that the two web sources produced very similar results (Pearson correlation coefficient: 0.96), although the correlation is much higher for level-2 (0.94) and level-1 (0.91) proverbs, than for level-0 (only 0.86).

**Table 1.** Total and average frequency of proverbs matched by the web browsers Google and Bing, and their sum (G+B), considering the entire sample (100 proverbs; 'All'), and by reference level ('L-2' to 'L-0').

| | Total | | | Average | | |
|---|---|---|---|---|---|---|
| **Level** | **Google** | **Bing** | **G+B** | **Google** | **Bing** | **G+B** |
| All | 2,594 | 7,143 | 9,143 | 26 | 71 | 97 |
| L-2 | 1,650 | 5,196 | 6,846 | 66 | 208 | 274 |
| L-1 | 833 | 1,834 | 2,667 | 33 | 73 | 107 |
| L-0 | 111 | 113 | 224 | 2 | 2 | 4 |

Table 1 shows the total and the average frequency of proverbs matched by the browsers, and the breakdown by reference level of lexical availability. It is noteworthy that Bing produced 2.7 times more matches than Google. When compared against the reference, the frequency values were slightly higher for the results from Google (Pearson: 0.73) than for those from Bing (0.69), while the sum of the frequency values from both browsers yield an intermediate value (0.70). This corresponds to a relatively high correlation.

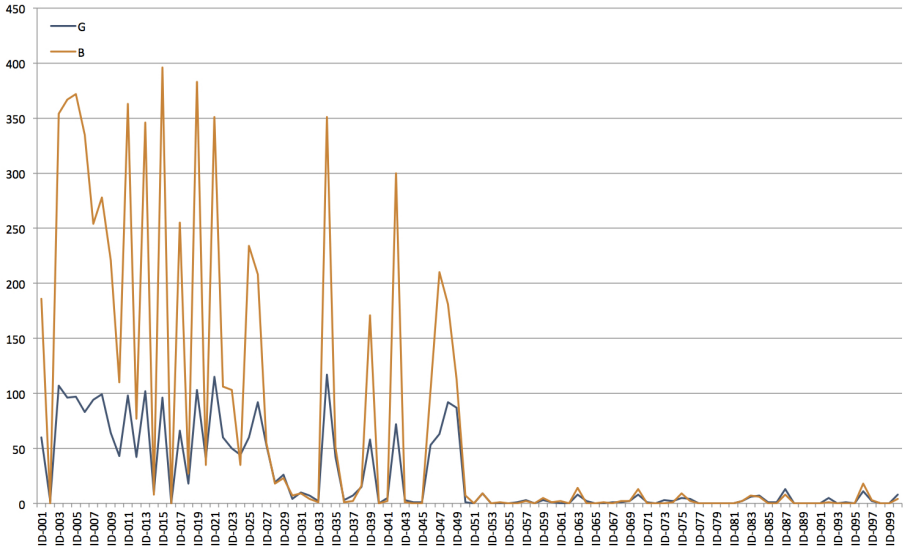


**Fig. 2.** Proverbs Frequency in two web browsers. Horizontal axis: the sample of 100 proverbs, identified by their ID code; from ID-001 to ID-025: reference level 2 (highly available); from ID-026 to ID-050: reference level 1 (moderately available); from ID- 051 to ID-100: reference level 0 (seldom available). Vertical axis: hit counts, exact match, top domain Portugal (.pt), language: Portuguese (Portugal); values: Google (G, retrieved on 2017/05/03) and Bing (B, retrieved on 2017/05/13).

## 4 Proverbs in *corpus*

Finally, the same proverbs were searched in the Cetem*Público corpus* [23][4]. This is a large-sized, publicly available *corpus* of journalistic text, collected from the online edition of the European Portuguese newspaper *Público*, and containing about 9,6 million words.

To process and query the *corpus*, the UNITEX linguistic development platform [24][5] was used. The *corpus* was processed using the European Portuguese language resources distributed with the system. The queries were carried out using finite-state transducers (FST) that are built using this platform formalism. These FST define a linguistic pattern to be matched and output the ID of the proverb corresponding to the matched string. Figures 3 and 4 show the FST used for querying the proverb *Santos de casa não fazem milagres* 'Home saints don't make miracles'.

[4] www.linguateca.pt/cetempublico
[5] http://unitexgramlab.org/

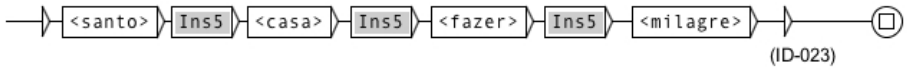


**Fig. 3.** FST built for querying the proverb *Santos de casa não fazem milagres* 'Home saints don't make miracles' in *corpora*: sequences of keywords. Words inside chevrons represent lemmas. Grey boxes *Ins5* are subgraphs for insertions of 0 up to 5 words.

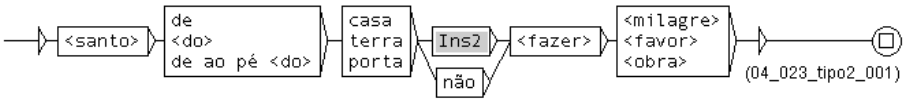


**Fig. 4.** FST built for querying the proverb *Santos de casa não fazem milagres* 'Home saints don't make miracles' in *corpora*: graph describing lexical variants. Words inside chevrons represent lemmas. Grey box *Ins2* is a subgraph for insertions of 0 up to 2 words.

Two methods of querying were used:

a) a set of FST describing the sequence of keywords (mainly nouns, verbs and adjectives) that characterise the proverb, allowing for a window of 0 to 5 words between them, and a small set of punctuation marks[6]: < ; , ( [ ) ] / and . . .>. This type of FST was automatically built from the database. Keywords are represented by their lemmas.
b) another set of FST, manually built to describe all the variants of a given proverb found in the database, or deemed as reasonably probable to occur, according to the vocabulary involved and the syntactic structure of the proverb. These FST are only available for the 50 proverbs from levels 2 and 1, since level 0 expressions were not considered sufficiently relevant to be represented in this way.

Table 2 shows the results from applying the two sets of graphs to the Cetem*Público corpus*. As expected, only some few instances (34) of the sampled proverbs were found in the *corpus*, corresponding to 13 different proverbs, from which 30 proverbs (instances) were matched by both methods. The keywords' graphs matched 40 instances (hence 10 false-positives), while the variants' graphs are very precise (no false-positive matches).

The (small) difference between the results from the variants' graphs and from the keywords' graphs (proverbs ID-026, ID-038 and ID-050) is due to the fact that the variants' graphs include lexical variants of the keywords, while the keywords' graphs only consider one of those lexical variants.

For example, for proverb:

ID-026 *O sol quando nasce é para todos*
'The sun when it is born is for everyone'

[6] These insertions are reported in the sub-graphs - the grey boxes in Figures 3 and 4.

**Table 2.** Sampled proverbs matched in the Cetem*Público corpus*. Column 'Var indicates the number of matches using the proverbs' variants FST, while column 'Key' corresponds to FST with the keywords. An approximate translation of the proverbs (or an equivalent expression) is provided.

| ID | Level | Proverb 'equivalent/translation' | Var | Key |
|---|---|---|---|---|
| ID-003 | 2 | *Depois da tempestade vem a bonança*<br>'After the storm comes the calm' | 2 | 2 |
| ID-013 | 2 | *Quem sabe sabe*<br>'Who knows, knows' | 4 | 4 |
| ID-015 | 2 | *Mais vale tarde do que nunca*<br>'Better late than never' | 1 | 1 |
| ID-019 | 2 | *O tempo voa*<br>'Time flies' | 1 | 1 |
| ID-021 | 2 | *O seguro morreu de velho*<br>'The careful man died of old age' | 5 | 5 |
| ID-022 | 2 | *A esperança é sempre a última a morrer*<br>'Hope is the last to die' | 3 | 3 |
| ID-023 | 2 | *Santos de casa não fazem milagres*<br>'Home saints don't make miracles' | 7 | 7 |
| ID-026 | 2 | *O sol quando nasce é para todos*<br>'The sun when it is born is for everyone' | 2 | 0 |
| ID-034 | 1 | *Nem só de pão vive o homem*<br>'Man does not live by bread alone' | 2 | 2 |
| ID-038 | 1 | *Um crime não justifica outro*<br>'One crime does not justify another' | 1 | 0 |
| ID-042 | 1 | *Quem avisa, amigo é*<br>'He who warns is a friend' | 1 | 1 |
| ID-049 | 1 | *O tempo é dinheiro*<br>'Time is money' | 4 | 4 |
| ID-050 | 1 | *Não degenera quem sai aos seus*<br>'He who resembles his own [people] does not degenerate' | 1 | 0 |
| | | **Total** | **34** | **30** |

the keywords' graph only considers the keywords *sol-nascer-ser-todos* 'sun-rise- be-everyone', but the corresponding variants' graph allows *brilhar* 'shine' instead of *ser* 'be'.

For proverb:

ID-038 *Um crime não justifica outro*
'One crime does not justify another'

the lexical variant *erro* 'mistake' is allowed but it was not a keyword.

Finally, for proverb:

ID-050 *Não degenera quem sai aos seus*
'He who resembles his own people does not degenerate'

the lexical variant *puxar* 'pull/take' was represented in the variants' graph, but not in the keywords.

Only level-2 and level-1 proverbs were matched, 8 and 5, respectively. Though the number of instances is low, the most frequently occurring, level-1 proverbs, namely,

ID-049 *Tempo é dinheiro* (4 matches)
'Time is money'

and (the fragment of) the biblical quote

ID-034 *Nem só de pão vive o homem* (2 matches)
'Man does not live by bread alone' (Mt 4:4)

may lead us to reclassify them as level-2.

Notice that the FST account for morphosyntactic variation of the proverbs' key-words. For example, the gender-number variation of *santo* 'saint', or number variation on *milagre* 'miracle', as well as subject-verb agreement, are all handled by the graphs, using the lexical resources available with the system:

… *Portanto, para a Qualidade Total***, santo de casa é quem faz milagres** ...
'Thus, for the Overall Quality, home saint (masc.-sg.) is the one that does miracles (pl.)'

... *E* **santa de casa não faz milagre** ...
'Home saint (fem.-sg.) doesn't do miracle'

… *contrariando a tese de que* **santo de casa não faz milagre** ...
'against the thesis that home saint (masc.-sg.) doesn't do miracle (sg.)'

The proverb:

ID-050 *Não degenera quem sai aos seus*
'He who resembles his own people does not degenerate'

is more often used with the subject in the canonic word order, v.g.

*Quem sai aos seus não degenera.*

This variant had been used in the survey because of the random selection method used for the sampling. The more common variant, showing the basic word order was also checked. However – and quite surprisingly, no match was found in the *corpus*.

Contrasting with the low frequencies observed in this *corpus*, all these proverbs are quite frequent in the web (see Section 3). In average, these proverbs occurred 78 times in (Google) and 233 times (in Bing) – standard deviation, approximately 37 and 146, respectively; and totalling 1,017 and 3,034 respectively.

In spite of the low frequency observed in the *corpus*, it seems possible to conclude that:

a) most probably due to the journalist nature of the *corpus*, the occurrence of proverbs is this type of text is scarce, as expected;
b) even so, the lexical availability level, manually assigned to the proverbs' sample, has been confirmed, since no level-0 expressions were found, and there are more instances of level-2 than of level-1 proverbs.

## 5 Conclusion and Future Work

This paper set out to establish the lexical availability of a large-sized database of with over 114,000 proverbs. A preliminary estimation, carried out by two annotators, was confirmed by and large using data obtained through a survey (735 participants) and through queries on two popular web browsers (Google and Bing). Results from queries over a large-sized *corpus* of journalistic text also confirmed the initial expectations that the use of this type of linguistic expressions is often limited by style conventions to oral/colloquial communicative contexts.

In view of the results, it is reasonable to extend the lexically available status to all the level-2 manually selected proverbs. These will constitute the main core of a lexicon of commonly used proverbs. Level-1 proverbs, in general, and certain difficult or interesting cases, both from level-2 and level-0, will have to be studied further. In some cases, only some variants of a given proverb should be assigned level-2 status, while the remainder variants may be attributed to level-1 (moderately available) or even level-0 (seldom available). In other (rarer) cases, level-0 proverbs (or variants) may have to be raised to level-1, too.

The 50 finite-state transducers already built will now be extended to the remainder of level-2 proverbs and, time allowing, to level-1, after careful revision of this list. Eventually, surveying the lexically availability of the remaining entries from level-1 will be useful.

Previous experiments using a Brazilian Portuguese database of approximately 3,500 proverbs (614 types or paremiological units) over a relatively large *corpus* of this language variety [25] showed that the approach of using just the proverb's keywords (surface forms) is a reasonably effective strategy for detecting proverbs in texts.

In this paper, however, the keywords in the FST were (for the most part) lemmatized, which broadened the search area of the queries in the *corpus*.

A similar experiment has already been carried out on a *corpus* of Portuguese textbooks [15], but due to technical shortcomings of the linguistic platform used, it was not

possible to build FST with lemmatized keywords for the entire database. As a consequence, only the keywords' surface forms were used, narrowing the queries' search space. Having narrowed down in this paper the set of lexically available (or, at least, moderately available) proverbs of European Portuguese, it should now be possible to produce a new, richer resource, equivalent to that of [25], improving the accuracy and recall of proverb identification in texts.

More importantly, with this paper, an empirically motivated list of lexically available proverb (and variants) has now been produced[7], which can be used in a reliable way to develop many types of applications, for example, for diagnosis/therapy of some speech disorders or for didactic games for language learning.

**Acknowledgements**

This work was partially supported by national funds through Fundação para a Ciência e Tecnologia (FCT) with reference UID/CEC/50021/2013.

**References**

1. Charteris-Black, J.: Proverbs in communication. Journal of Multilingual and Multicultural Development, 16:4, 259–268 (1995).
2. Mieder, W.: Proverbs – A Handbook. Greenwood Press, London (2004).
3. Rezaei, A.: Rhetorical Function of Proverbs Based on Literary Genre. Procedia – Social and Behavioral Sciences 47, Elsevier Ltd. 1103–1108 (2012).
4. Meira, A.: Casa de ferreiro, espeto de pau: uma análise das relações retóricas a partir do uso dos provérbios como estratégia argumentativa em textos da internet. (Ph.D. thesis), Faculdade de Letras da UFMG, Belo Horizonte (2015).
5. Hrisztova-Gotthardt, H., Varga, M. (eds.): Introduction to Paremiology: A Comprehensive Guide to Proverb Studies. DeGryuter, Berlin (2015).
6. Council of Europe.: Common European Framework of Reference for Languages: Learning, Teaching. Council of Europe (2001).
7. Chaika, E.: Linguistics, Pragmatics and Psychotherapy – A Guide for Therapists. Whurr Publishers, London and Philadelphia (2000).
8. Arif, M., Abdullah, I.: The impact of output communication on EFL learners' metaphor second language acquisition. Social Sciences (Pakistan) 11.9 1940–1947 (2016).
9. Salbego, N., Osborne, D.: Schema activation through pre-reading activities: teaching proverbs in L2. BELT– Brazilian English Language Teaching Journal 7.2, 175–188 (2016).
10. Gorham, D.: A proverb test for clinical and experimental use. Psychological Reports, 1, 1–12 (1956).
11. Benton, A.: Differential behavioral effects in frontal lobe disease. Neuropsychologia 6, no. 1, 53–60 (1968).
12. Gibbs, R., Beitel, D.: What proverb understanding reveals about how people think. Psychological Bulletin 118.1, 133–154 (1995).

[7] This list is available to the scientific community at: https://www.researchgate.net/publication/319465467_List_of_100_proverbs_annotated_with_lexical_availability (DOI: 10.13140/RG.2.2.30110.64326).

13. Siqueira, M., Marques, D., Gibbs, R. Jr. Metaphor-related figurative language comprehension in clinical populations: a critical review. Scripta 20 (40), 36–60 (2016).
14. Vas, A., Spence, J., Eschler, B., Chapman, S.: Sensitivity and specificity of abstraction using gist reasoning measure in adults with traumatic brain injury. Journal of Applied Biobehavioral Research 21.4, 216–224 (2016).
15. Reis, S., Baptista, J. O uso de provérbios no ensino do Português [The use of proverbs in the teaching of Portuguese]. Proceedings of the 10th Interdisciplinary Colloquium on Proverbs, Tavira, Portugal, November 6–13, 2016 (in print).
16. Reis, S., Baptista, J.: O provérbio como estímulo num terapeuta virtual [Proverbs as a stymulus of a virtual therapist]. 6th Simpósio Mundial de Estudos sobre o Português (SIMELP), Simpósio 77, A Importância da Aprendizagem Lexical, Santarém, Escola Superior de Educação, Instituto Politécnico de Santarém (2017, accepted for publication).
17. Reis, S., Baptista, J. 2016b. 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, 5th-6th May, University of Algarve, Faro, Portugal, 427–446.
18. Abad, A., Pompili, A., Costa, A., Trancoso, I., Fonseca, J., Leal, G., Farrajota, L., Martins, I. Automatic word naming recognition for an on-line aphasia treatment system. Computer Speech and Language, Elsevier, 27 (6), 1235–1248 (2013).
19. Coxhead, A., Nation, P., Sim, D. Measuring the vocabulary size of native speakers of English in New Zealand Secondary schools. New Zealand Journal of Education Studies (2015).
20. Martins, E. Manual de redação e estilo. 3rd ed., São Paulo, O Estado de S. Paulo (1997).
21. Chacoto, Lucília. Estudo e Formalização das Propriedades Léxico-Sintácticas das Expressões Fixas Proverbiais, (Master thesis), Faculdade de Letras da Universidade de Lisboa, Lisboa (1994).
22. Reis, S., Baptista, J.: Portuguese Proverbs: Types and Variants. In: Corpas Pastor, G. (ed.). Computerised and Corpus-based Approaches to Phraseology: Monolingual and Multilingual Perspectives, 208–217. Editions Tradulex, Geneva (2016).
23. Santos, D., Rocha, P.: Evaluating CETEMPúblico, a free resource for Portuguese. In Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (Toulouse, 9–11 de julho de 2001), pp. 442–449 (2001).
24. Paumier, S.: Unitex 3.1 User Manual. Université de Paris-Est/Marne-la-Vallée – Institut Gaspard Monge, Noisy-Champs (2016).
25. Rassi, A., Baptista, J., Vale, O.: Automatic Detection of Proverbs and their Variants. In: Pereira, M., Leal, J., Simões, A. (eds.): Proceedings of the Symposium on Languages, Applications and Technologies (SLATE'14), Leibniz (Germany), pp. 235–249. Schloss Dagstuhl - Leibniz-Zentrum fur Informatik, Dagstuhl Publishing (2014).