LET'S PLAY WITH PROVERBS?
NLP tools and resources for iCALL applications around proverbs for PF

Sónia Reis [1] and *Jorge Baptista* [1,2]
[1] Universidade do Algarve, Faculdade de Ciências Humanas e Sociais
[2] INESC-ID Lisboa, Spoken Language Laboratory
reis.soniamm@gmail.com, jbaptis@ualg.pt

ABSTRACT
Proverbs are an important form of cultural expression of a society and are related to various areas of knowledge and human experience (González Rey, 2002). While linguistic elements in widespread use, proverbs are very rich structures both from a cultural and from a linguistic point of view and can therefore contribute significantly to the teaching of languages, both native and foreign (Council of Europe, 2001). However, though there are extensive collections of Portuguese proverbs with tens of thousands of forms and its variants (Reis, *in preparation*), its automatic identification in texts is quite difficult, given its formal variation, both lexical and syntactic (Chacoto, 1994). Nevertheless, using real examples, where proverbs are used in a natural or spontaneous discourse context, is a more natural way to learn and teach the complex conditions and communicative situations that determine the use and meaning of these expressions. On the other hand, frequency indices associated with proverbs and its variants would allow one to select the most common expressions. These are precisely the most interesting forms from the point of view of their teaching/learning and could serve as a basis for the construction of educational games, particularly for learning Portuguese autonomously as a foreign language (PFL) assisted by computer. To make this possible, it is necessary, first of all, be able to recognize the occurrence of proverbs in the texts (Rassi *et al.* 2014), including the instances where these expressions are presented in a truncated or creatively modified form, for example, to better suit the communicative situation or to produce new and more expressive meanings. In this paper, we present an on-going project, which aims at automatic identification of proverbs in texts. In this interdisciplinary study, we combine natural language processing tools with questionnaires construction techniques for teaching purposes (Hoshino and Nakagawa 2005, Correia *et al.* 2010). This is illustrated here with different sets of formats that can be built based on the knowledge of the form and variation of proverbs, as well as their frequency in *corpora*.

Keywords: Portuguese proverbs, Intelligent Computer-Assisted Language Learning (iCALL), Natural Language Processing (NLP), didactic gaming
JEL Classification: Z00

1. INTRODUCTION
Proverbs are an important form of cultural expression of a society and are related to various areas of knowledge and human experience (González Rey, 2002). This type of expressions has a relatively frozen structure, usually formed by short phrases, and often composed of two (sometimes three, or more) parts. Also, it is common to find alliterations and rhymes in proverbs. These are all mnemonic features denoting their transmission process, which is predominantly oral, and they generic value as words of wisdom (maxim, precept, etc.). In spite of its frozen characteristics, because of their oral tradition, they also present a relatively wide spectrum of lexical and syntactic variation (Chacoto, 1994). This formal variation makes it particularly difficult the automatic identification of proverbs in texts, which constitutes a challenge to Natural Language Processing (NLP). However, their automatic detection and delimitation not only would be relevant for natural language understanding and and automatic discourse analysis, as it could also prove to be extremely useful for several NLP applications (Witten *et al*. 2011), such as machine translation, or even to language learning and teaching. It is on this latter field of application that this paper will focus.

In order to use proverbs as linguistic material for language learning and teaching, it is necessary, in the first place, to be able to retrieve them from the texts where they occur, and especially those

that occur more frequently and in a larger set of communicative contexts. In order to be able to produce such careful selection for pedagogic purposes, it is important that frequency indices be associated to the proverbs and their variants.

In this paper, we focus on the types of activities (didactic exercises) that it is possible to develop around proverbs in the framework of computer-assisted language learning (CALL). More precisely, we adopt the CALL perspective inspiring the REAP.PT platform, presenting a brief overview of its most salient features, in order to draw a set of exercises highlighting the linguistic resources and NLP tools required to build such complex objects. Our main goal is to set up a roadmap in view of developing those applications, which we expect to be able to build, in the future.

The paper is organized as follows: In 1.1. some general issues about proverbs and their role in language learning are addressed, particularly in an iCALL context for Portuguese as a Foreign Language (PFL). In 2, existing on-line exercises are presented and commented. In 3, the main NLP tools and resources for NLP helpful for building exercises on proverbs are briefly presented. In 4, the *corpus* of proverbs and its construction is briefly described. In 5, some remarks are made on the automatic identification of proverbs in texts, and in 6 we present several suggestions regarding the automatic generation of exercises with proverbs using the NLP tools and language resources mentioned in the preceding sections. Finally, in 7, the paper is summarized and future work is considered.

1.1 General Issues on teaching proverbs in a iCALL-PFL environment

While linguistic elements in widespread use, proverbs are very rich structures both from a cultural and from a linguistic point of view and can therefore contribute significantly to the teaching of languages, both native and foreign. According to the *Common European Framework of Reference for Languages* (Council of Europe, 2001), the communicative language competence comprises three components: (a) linguistic competences; (b) sociolinguistic competences and (c) pragmatic competences. *Lexical competence*, an integral part of linguistic competences, concerns the "knowledge of, and ability to use, the vocabulary of a language" (*idem*: 110) and it includes the knowledge of both grammatical and lexical elements, including proverbs and other fixed expressions. "Sociolinguistic competence is concerned with the knowledge and skills required to deal with the social dimension of language use" (*idem*: 118), and also includes proverbs since they are "fixed formulae which both incorporate and reinforce common attitudes, make a significant contribution to popular culture" (*idem*: 120).

Information technologies have progressively acquired a more important role in teaching, and particularly in language teaching. Computer-assisted language learning platforms allow students to have a more active role in the learning process, enabling them to exercise in an autonomous way their skills and, at the same time, providing constructive feedback on their progress. They are also an important tool to monitor students' performance and progress, giving the teacher the means to target, in a more personalized way, his/her intervention.

The REAP.PT[1] is a computer-assisted language-learning platform developed for the teaching and learning of Portuguese. It is a powerful aid to the learning process, both for native speakers and for students of Portuguese as a foreign language, and it also constitutes a valuable auxiliary tool to the teacher. The system is specially aimed at the learning of vocabulary and it is based on the intensive use of Natural Language Processing (NLP) tools and techniques in order to build automatically a large variety of exercises, drawn from *corpora* of real texts, with their contents adapted to the topic preferences of the students. Many exercises consist of multiple-choice questions, in which a target expression (in the question) is presented (in the set of possible answers) along with a set of *distractors* (a.k.a. *foils*), usually no more than three, each one selected based on a set of linguistically motivated criteria. In this process, the sentences with the target expression are automatically retrieved from *corpora*, while the set of distractors is automatically built using NLP tools and techniques, using with linguistic resources adapted to that purpose. Furthermore, the REAP.PT system is conceived in order to be able to built a dynamic model of the student, keeping track of his/her progress and al-

[1] http://www.l2f.inesc-id.pt/wiki/index.php/REAP.PT (last access: March 31, 2016; all the remaining URL mentioned in this paper have been check on this date).

lowing the teacher to monitor it. Finally, the system, and most prominently, the system automatically correct the exercises and, as much as possible, produces some positive feedback about the linguistic aspect at stake, or alerting the student for the nature of incorrect answer (Pellegrini *et al*. 2012).

This paper intends to lay down the basis for the construction of didactic games with proverbs, in view of integrating them into the REAP.PT platform. With this games involving proverbs, students will be able to identify and use this type of culturally rich trove of expressions, hone their language skills, and evolve their vocabulary competence. To this end, we will first present an overview of the type of exercises involving proverbs that can be found in different websites and computer-assisted language learning platforms, highlighting the challenges each type poses to its construction under the REAP.PT approach to language gamming, focusing on the type of linguistic resources and natural language processing techniques required to put them into practice.

2. SURVEY OF LANGUAGE-RELATED EXERCISES WITH PROVERBS

In this section, we present a brief survey of existing exercises built around proverbs or focusing on the learning form and meaning, including certain CALL platforms, especially in view of learning Portuguese as a foreign language.

A large part of the games and exercises on proverbs made available in several websites adopt a format whose goal is to complete the expression, either half of the proverb is provided so that the user is supposed to know and complete the other half, or one or more words had been removed and the uses had to fill in the corresponding blank spaces. The purpose of these games and exercises, besides promoting the better knowledge of the proverbial stock of the language, can sometimes be inferred as targeting several language skills. In the following lines, we present some of the most interesting of these websites' games and exercises.

The game *Finish the Proverb*[2] consists, basically, in a multiple-choice answering task, completing the proverbs from which one or more words have been removed. Several exercises (10) are presented at a time, each featuring a proverb that must be completed, and a set of four answers is provided, one being the correct answer and the other three the distractors (or foils). The exercise is automatically corrected but only after the entire set of questions has been answered (and not one at a time). In the correction, the user gets to see the correct answer for each question. The questions form a closed set and, as far as we could find, they have been manually constructed. One of the interesting aspects of this site is the fact that, in the feedback, the user is shown the percentage of players that got that answer right (but not the absolute number!). This could be an indicator of the frequency of the lexical availability of those proverbs. The meaning of each proverb is also explained.

The set of proverbs produced in this game aim at different linguistic competences from the player. In the next table, we try, in a non-exhaustive way, to identify those language skills. These linguistic aspects concern both the target word and the distractors, as well as the interplay between them, so that, on one hand, their decoding is not always easy, and on the other hand, it is sometimes difficult to grasp exactly the concepts at stake, often more than once at the same time. Hence, these are complex exercises, plainly justifying the human effort put into their construction, and highlighting the challenge that their automation involves. Table 1 provides a detailed analysis of the type of relations the distractors hold with the target word, for each proverb.

Several linguistics phenomena are dealt with in these proverbs, as the comments provided for each distractor and highlight, which are associated either to the structure or to the meaning of these expressions. In other words, distractors are never random, each has been produced with a purpose in mind. Each aspect that a distractor involves provides the student with the opportunity to test his/her linguistic competences, and often several of them at the same time, particularly his/her lexical and semantic knowledge. All this is done in a non-explicit way but still very much "applied" situation, while the gaming context, with the scoring or agonistic situation, entails a more appealing stimulus. Among the relevant phenomena, the most prominent are: (i) the semantic relations between word senses (synonymy and antonymy; holonymy and meronymy; hypernymy and hyponymy;

[2] www.123facts.com/play-quiz/Finish-The-Proverb-5244.html

cause-effect) (ii) symbolic values invested in lexical items, having to do with culture and world-knowledge; (iii) collocational nature of word combinations[3]; and (iv) formal aspects such as rhyme, the two- or threefold partition of proverbs (*hemistichs*), and paronymy.

Table 1: Linguistic categories/phenomena and language skills mobilized in the task of completing proverbs

Question	Proverb & Answers	Linguistic phenomena
1	**Hatred is as blind as ...** **(D1)** A mole **(CA)** Love **(D2)** Darkness **(D3)** None of these	**Correct answer**: antonymy *hatred*/*love* and a semantic relation with the proverb *Love is blind* ***Distractor* 1**: relation with the member of a set (animal/mamiffer) and analogy with the frozen idiomatic adverb *to blind as a bat* ***Distractor* 2**: semantic association between *blindness* and *darkness* ***Distractor* 3**: something else (unknown target)
2	**Blood is thicker than ...** **(D1)** Wine **(CA)** Water **(D2)** Tears **(D3)** Sweat	**Correct answer**: (unmotivated, idiomatic) ***Distractor* 1**: semantic but culturally-founded relation (in Christian tradition, the eucharistic wine is associated with blood) ***Distractor* 2**: member of the set of objects to which *blood* can be associated, and which could be subsumed under the term *bodily fluids* ***Distractor* 3**: same as above
3	**A stitch in time saves ...** **(D1)** Thousands **(D2)** A dime **(CA)** Nine **(D3)** Everybody	**Correct answer**: rhyme *time/nine* ***Distractor* 1**: both *a* and the numeral *thousands* are determiners *stitch*; the numeral is hyperbolic (= *many stiches*) and it is the opposite of *a* (single) *stich* ***Distractor* 2**: rhyme *time/nine* ***Distractor* 3**: the distractor is an indefinite pronoun distributionally adequate as direct object of the verb, thus producing the generic reading typical of proverbs
4	**Tall oaks grow from ...** **(D1)** Strong earth **(D2)** Nutritious soil **(D3)** Small saplings **(CA)** Little acorns	**Correct answer**: antonymy *tall*/*little* and semantic relation between *oak* and *acorn*; ***Distractor* 1**: the noun *earth* is semantically motivated as locative complement of grow from ~ and is quasi-synonymous of *soil*; *strong* is a collocational modifier of *soil*; ***Distractor* 2**: *idem*; *nutritious* is also a collocational modifier of *soil*; ***Distractor* 3**: antonymy *tall*/*small* and semantic relation between *oak* a (tree species) e *sapling* (a development stage of a tree).
5	**The ... wounds more than a lance** **(D1)** Hand **(CA)** Tongue **(D2)** Word **(D3)** Evil eye	**Correct answer**: (unmotivated, idiomatic): the body-part *tongue* connotes with *language*; ***Distractor* 1**: metonymy between *hand* and the agent (human) subject of *wound*; contrast between agent and the instrument (*lance*) ***Distractor* 2**: *word* also connotes *language*; ***Distractor* 3**: the compound *evil eye* is formed with a body-part noun (like CA *tongue* and D1 *hand*); contrast between a physical instrument (lance) and a symbolic one (evil eye).

In this same website, there are other quizzes involving proverbs. *Finish the Proverb Vol.2* [4] is quite similar to the game presented above, but it also includes some questions whose goal is to identify the meaning of the proverb from four possible answers. This is also the goal of the quiz *Guess the Proverb*[5], though this also includes questions aimed at identifying the sentence expressing the opposite meaning of the target proverb. Finally, in *Complete the Proverb/Idiom*[6], is similar to the exercise we presented at length above: ten well-known idioms and proverbs are produced with a missing word that the player must complete from a set of four possible answers. However, in this case, the feedback indicates that the players guessed 100% of answers correctly, which may indicate that this quiz is too easy, and may not be challenging enough for the user. To conclude, it is worth mentioning that the site allows players to rate, bookmark and recommend it to other users.

Table 1 (cont.)

[3] The term collocation is used here in the sense of Mel'čuk & Polguère (2007): a word (the *collocate*) that establishes a well defined semantic relation (*lexical function*) with another word (the *base*); the choice of the collocate is lexically determined, usually idiosyncratic. Lexical functions are a finite, though large, set of semantic and syntactic relations such as qualitative-positive assessment BONUS ('good'), which is the case here: given the base *soil* a finite set of adjectives exist to express its quality (in view of its use) -- *rich*, *nutritious*, etc.

[4] http://www.123facts.com/play-quiz/Finish-The-Proverb-Vol-2-5305.html

[5] http://www.123facts.com/play-quiz/Guess-The-Proverb-5297.html

[6] http://www.123facts.com/play-quiz/Complete-the-Proverb-Idiom-7046.html

Question	Proverb & Answers	Linguistic phenomena
6	**Two wrongs don't make ...** **(CA)** A right **(D1)** It easier **(D2)** A fight **(D3)** The world a better place	**Correct answer**: antonymy *wrong*/*right* ***Distractor* 1**: though the sentence is acceptable, the relation with the target word is unclear ***Distractor* 2**: Rima *fight* / *right* (CA) ***Distractor* 3**: though the sentence is acceptable, the relation with the target word is unclear
7	**Make a silk purse out of a ...** **(D1)** Silk dress **(D2)** Woolen sock **(CA)** Sow's ear **(D3)** Cow's ear	**Correct answer**: (unmotivated, idiomatic): opposition between a noble/expensive raw material (*silk*) and a cheap/gross one, with a negative connotation (*sow's ear*); ***Distractor* 1**: the same material adjective (*silk*) and a clothing item (*dress*), which can be associated with *purse*; ***Distractor* 2**: another material adjective (*woolen*) outro adjetivo de matéria (*woolen*) and a clothing item (*sock*) ***Distractor* 3**: paronymy (*sow*/*cow*) with the target word, both related to *livestock* (hypernym); the same body-part noun (*ear*)
8	**A loaded ... makes no noise** **(D1)** Horse **(D2)** Gun **(D3)** Train **(CA)** Wagon	**Correct answer**: collocation between *loaded* and *wagon*; appeal to world knowledge, equating *make noise* with *speak loudly* and *loaded* with *richness*; ***Distractor* 1**: collocation between *loaded* and *horse*; ***Distractor* 2**: collocation between *loaded* and *gun*; ***Distractor* 3**: collocation between *loaded* and *train*; meronymy relation between *wagon* and *train*;
9	**A burnt child dreads the ...** **(D1)** Match **(D2)** Stove **(D3)** Air conditioner **(CA)** Fire	**Correct answer**: semantic relation (cause-effect) between *fire* and *burn*/*burnt*; ***Distractor* 1**: semantic relation between *fire* and *match* (instrument); ***Distractor* 2**: semantic relation between *fire* and *oven* (object-tool to generate heat) and, indirectly, a semantic relation (cause-effect) between *oven/heat* and *burn*/*burnt*; ***Distractor* 3**: semantic relation between *air conditioner* and the distractor *stove* (both subsumed under the hypernym *domestic appliances*), also involving the concept of *temperature*.
10	**Learn to walk before you ...** **(D1)** Crawl **(D2)** Speak **(CA)** Run **(D3)** Leap	**Correct answer**: *walk* and *run* are both motion verbs but run implies more speed, thus there is a gradation that motivates the order of these verbs in the proverb; ***Distractor* 1**: the relation between *crawl* and *walk* is similar as in *walk*/*run* (CA) but the gradation is inverted; ***Distractor* 2**: *walk* and *speak* represent two successive stages in child development; ***Distractor* 3**: the relation between *leap* and *walk* is similar (motion verbs) but the comparing parameter is not *speed* but rather *complexity* or *distance*.

CA= Correct answer; D1= Distractor 1; D2= Distractor 2; D3= Distractor 3; the order of the answers, which may also be relevant for the task, is the same as shown in the website. Apparently, this order is constant.

The *Centro Virtual Camões*[7] is one of the Portuguese websites that produces a set of language-related games/quizzes. The *Jogo da Lusofonia*[8] is a game about the Portuguese-speaking countries, providing 256 multiple-choice questions, drawn from different topics, including some proverbs and idioms. The player, who can choose to play 5, 10 or 15 questions at a time, can customize each run. The user also has 3 opportunities to ask for help. Answers consist of 3 choices and are timed (30 seconds each), which adds to the interest of the game. The game can be played alone or with a team, and with a number of other playing partners (or teams) as adversaries. The goal of the game is to achieve the maximal number of correct answers in the least time.

As far as the proverbs are concerned, most questions in this game consist in completing the proverb or choosing its correct meaning. The questions have been manually built and the authors have also produced the correction. In the feedback, along with the correct answer, the meaning of the proverb is also explained. Due to the overall goal of the game, some of the proverbs found in the game may not be part of the common stock of Portuguese-speaking countries (at least they were absent from the (European) Portuguese collections of proverbs known to us). For example, the Guinean proverb: *Os dentes do elefante não o cansam* 'The teeth of the elephant do not tire it' is explained as follows (our translation): "In Guinea-Bissau, the elephant is a symbol for parents, who, in spite of the material difficulties, are tireless in caring and raising their children". However, in our European cul-

[7] http://cvc.instituto-camoes.pt/
[8] http://cvc.instituto-camoes.pt/aprender-portugues/a-brincar/jogo-da-lusofonia.html

ture, elephants do not have the same connotations and shown by the five proverbs[9] we find in our *corpus*, where this animal is mentioned:

(a) *A sebe dura três anos, o cão três sebes, o cavalo três cães, o homem três cavalos, o corvo três homens e o elefante três corvos* (The hedge lasts three years, the dog three hedges, the horse three dogs, three horses a man, the crow three men and the elephant three crows)
(b) *Quando dois elefantes lutam a erva fica espezinhada* (When two elephants fight the grass gets trampled.)
(c) *Com um cabelo de mulher, pode amarrar-se um elefante* (With a woman's string of hair one can tie up an elephant)
(d) *Elefante grandorro de pequerruchinho cresce* (A big elephant grows from a little one)
(e) *O teu inimigo é pequeno como uma formiga, mas guarda-te dele como se fosse um elefante* (Your enemy is small as an ant, but beware yourself from him as if he were an elephant)

In these proverbs, the elephant can connote with: longevity (a): unwanted/collateral effect of an action, due to the size (b); the overpowering feminine influence over men: contrasting size/strength of a string of hair and an elephant (c); the size/strength of an adult animal can not be guessed from its size as a youngling: contrast between the size of a young and a grown elephant (d); prudence towards an enemy irrespective of its size: contrast between the size of an ant and an elephant.

Besides these proverbs, we also found in the game the idiom *apanhar pulgas e deixar passar elefantes* (lit: catch fleas and let elephants pass, 'to be mindful of small details but ignore the larger mistakes), which we do not treat as proverb because of its subject being a distributionally free syntactic slot.

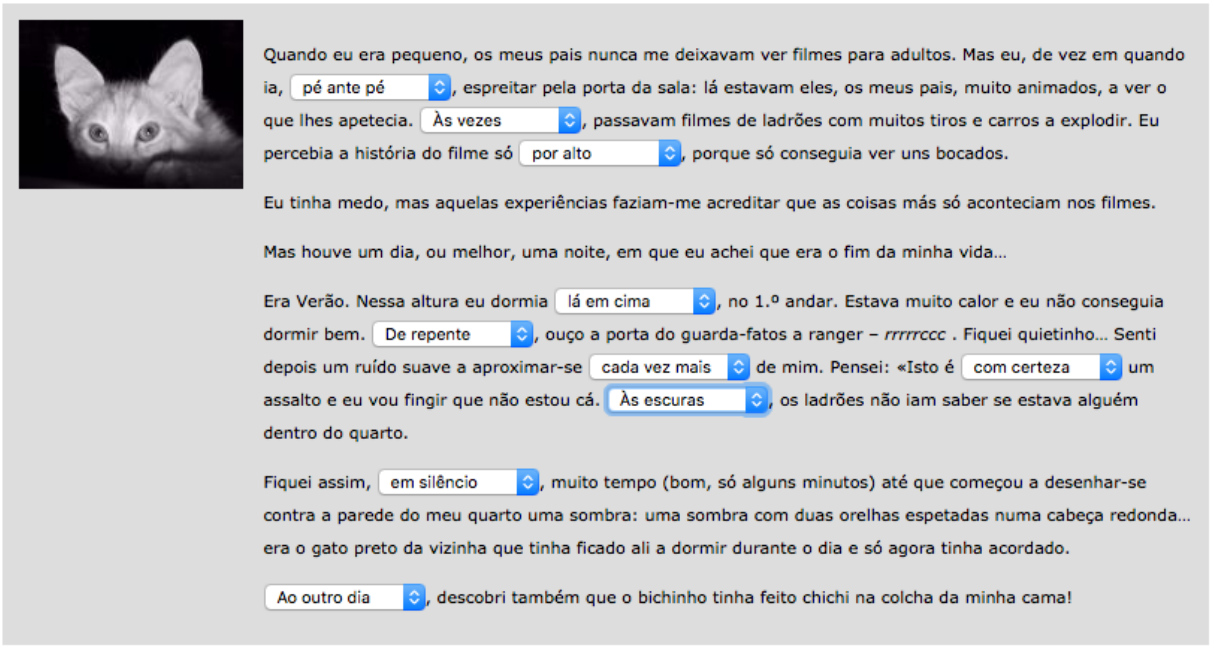
In the *Jogo da Lusofonia*, it was not possible to determine how many questions were there about proverbs; still, whenever we played it, we found several questions about proverbs and idioms. Another feature of the game is the feature of presenting on the top-left side the national flag of the Portuguese-speaking country the question is about. Notice, however, that many proverbs from the Brazilian Portuguese also belong to the European Portuguese stock, even if such fact is not signaled by the game. It was not possible to ascertain why was such or such proverb assigned to one the varieties, instead of the other one. Nevertheless, when the correct answer is explain, one often finds the indication that there is a corresponding proverb in more than one of the Portuguese-speaking countries. Fig. 1 is an example of such type of indication. In this case, the proverb appearing in the question is exclusive of the Brazilian variety of Portuguese, which is hinted by the name of the animal, the *urubu*, a vulture-like species that is endogenous of South America. Furthermore, in this case, the proverb is not exactly explained; instead, a Portuguese equivalent (and yet another, related or almost equivalent) proverb is provided.

The *Ciberescola da Língua Portuguesa*[10] is language learning platform that provides support for Portuguese as a Foreign Language (PFL) by making available several interactive language resources and online courses, aimed both at teachers and students, since they can be used within the classroom context, in a collaborative process, as well as a self-study tool. This platform covers all level of study, from the 5th to the 12th grade, and it also includes training material for the A2 and B2 levels of PFL. Students can choose their topics of interest (music and other arts, sport, traveling, folklore, etc.) and practice Portuguese in a funny and autonomous way, with exercises aimed at different language competences, namely, the training of oral and written understanding. To the best of our knowledge, no games or exercises involving proverbs have been produced yet for this platform.

[9] In the *corpus* there are 16 instances of these proverbs with *elefante* 'elephant', distributed by several variants and repeated versions, that is, occurring simultaneously in more than one collection.
[10] http://www.ciberescola.com/

Fig. 1: Jogo da Lusofonia

Fig. 2: Exercise on frozen expressions ("expressões fixas")



Concerning idiomatic expressions, only a fill-in-the-blank exercise with frozen expressions (and collocations) has been found (Fig. 2). The goal of this exercise, marked as "difficult", is to fill-in the ten blank spaces, and for each one 10 solutions are presented. After all the spaces have been completed, the form is submitted and the player's score is provided: this is the only feedback the application produces, the number of correct/incorrect answers, but not the solution to the exercise. In some cases, the answer-choices were shown in uppercase, thus hinting on the correct answer whenever that slot was in the beginning of a sentence. Most of these idioms are compound/frozen adverbs (Gross, 1996; Ranchhod, 2003; Palma, 2009): *com certeza* (= *certamente*), *em silêncio* (= *silenciosamente*).

In this exercise, those adverbs express the circumstances of *time* (frequency: *às vezes* 'sometimes', date-relative: *ao outro dia* 'the next day'), *manner* (most of them: *pé ante pé* 'on tiptoes', *por alto* 'vaguely', *etc*.), *place* (*lá em cima* 'upstairs') and *attitude* (Molinier and Levrier, 2000; *com certeza* 'certainly'). One of the issues with the exercise comes from the fact that for certain slots more than one possible solution if produced. For example: *Fiquei assim, em silêncio/lá em cima/às escuras, muito tempo* 'I stayed like that, in silence/upstairs/in the dark'; *Nessa altura eu dormia lá em cima/em silêncio, no 1º andar* 'At that time, I used to sleep upstairs/in silence'.

The *Observatório da Língua Portuguesa*[11] is an association whose main goal is to promote the Portuguese Language. Its webpage presents several resources to help Portuguese learners, including a weekly podcast[12] for teaching proverbs and idioms to students from the B2 and C2 levels. To date, only 39 idiomatic expressions and proverbs of European Portuguese, a insignificant number in view of the size of the lexicon, are explained with some degree of detail, along with an mp3 downloadable audio file and the corresponding text. In some cases, different expressions are compared, for example, to clarify the distinction between the use of the proverbs *Filho de peixe sabe nadar* 'The son of a fish knows how to swim' and *Tal pai, tal filho* 'Like father, like son', or relating them with idioms like *ser a cara chapada de alguém* 'be the spitting image of smb' or *seguir as pegadas de alguém* 'follow the footsteps of smb.' Examples of the use of proverbs are also produced, including some cases of creative adaptation of a proverb to a situation, like the use of feminine equivalent forms in *Tal mãe, tal filha*, to adjust the proverb to female characters. Notice, however, that all these are fabricated examples, and not spontaneous utterances retrieved from *corpora*. This website does not feature any game/exercise involving proverbs, so its usefulness is somewhat limited.

3. LANGUAGE RESOURCES AND NLP TOOLS FOR SEMANTIC PROCESSING

In this section, we present several language resources and NLP tools that are already available for processing Portuguese texts, which may be useful to produce didactic games for learning the proverbs of the language. These resources consist, basically, of a vocabulary where words are explicitly related by a set of semantic relations. For lack of a more consensual term, we call them all *ontologies*.

The *WordNet.PT*[13] (Marrafa, 2001) is a database developed at the Centro de Linguística of Universidade de Lisboa in the same methodological and conceptual framework as the English WordNet (Fellbaum, 1998)[14]. It is a language resource containing 19,000 expressions from European Portuguese, from different parts-of-speech and several semantic domains, structured in semantically homogeneous sets (or *synsets*) and the semantic relations between them, such as meronymy (or part-whole relations), synonymy, antonymy, class, event participants and structure. This data can be used in several areas of Computational Linguistics and Language Engineering. More recently, the Brazilian counterpart, *WordNet.Br*[15] (Dias-da-Silva, 2013), includes 5,860 verbs structured in 3,713 synsets.

The *Linguateca* project website[16] also distributes several semantic language resources. *PAPEL*[17] (Oliveira *et al*. 2008) is a freely downloadable lexical resource that consists of 102,000 different words and 191,000 semantic relations. These were semi-automatically extracted from the *Dicionário da Língua Portuguesa*, edited by Porto Editora, and include 83,000 (approx. 44%) synonymy word pairs, and (approx. 26%) hyponymy relations.

[11] http://observatorio-lp.sapo.pt/pt
[12] http://sayitinportuguese.pt/podcasts/
[13] http://www.clul.ul.pt/clg/wordnetpt/index.html
[14] http://wordnet.princeton.edu/
[15] https://en.wikipedia.org/wiki/WordNet
[16] http://www.linguateca.pt
[17] www.linguateca.pt/PAPEL

Table 2: Semantically related entries for the keyword *pássaro* 'bird'.
Proverb: *Mais vale um pássaro na mão que dois a voar* 'Better a bird in the hand than two flying'

Language Resource	*Pássaro* 'bird' [semantic relation]: word(lemma)+
WordNet.PT	[quasi-synonym]: *piu-piu* [agent/cause-result]: *chilreio* [co-related with]: *áugure* [is a member of] *bando* [hyponym (type of)]: *ave* [hypernym (supertype of)]: *andorinha, arara, beija-flor, bico-de-lacre, canário, catatua, corvo , cotovia, cuco, estorninho, gralha, melro, papagaio, pardal, pega, periquito, pica-pau, pintassilgo, pombo, rola, rouxinol, ta-rambola, tentilhão, tordo, tucano*
PAPEL[18]	[synonym_N_of]: *cou-cou*, *melriacho*, *pássaro-cou-cou*, *pica-rei* [hyperonym]: *barrete*, *lavandisca*, *pardal*, *petinha*, *poupa*, ... (172 entries) [is a member of]: *ordem_de_pássaro* [made_with]: *alçapão*, *aramenha*, *arapuca*, *armação*, *boiz*, ... (14 entries)
Onto.PT	[synonym]: *ave, passarinho, folecha*
MWN.PT	[synonym]: *ave* [is a member of]: *bando, ave* [hiponym (is a kind of)]: *vertebrado*
PULO	[synonym][19]: *ave*
TEP 2.0	not found

Onto.PT[20] (Oliveira & Gomes, 2014) aims at building a lexical ontology for Portuguese, with information extracted from six different resources (dictionaries, thesauri, encyclopedias as well as from other ontologies and *corpora*[21]). It is also a freely downloadable ontology, and contains 156 lexical forms, organized in 117,000 *synsets* and more than 173,000 tuples and their semantic relations.

The *MNW.PT*[22]- MultiwordNet of Portuguese (Pianta *et al*. 2002) covers manually allegedly validated 17,200 synsets, connected by hyponymy and hyperonymy relations, concerning over 21,000 word senses/forms and 16,000 lemmas of European and Brazilian Portuguese. Is also includes sub-ontologies for the categories Person, Organization, Event, Location and ArtWorks, and 98 core base concepts, suggested by the Global Wordnet Association, along with 164 concepts defined by the EuroWordNet project.

PULO[23] (Simões & Guinovart, 2014) is a Portuguese Unified Lexical Ontology and works as a wordnet compatible with wordnets available for other languages (English, Galician, Basque, Spanish and Catalan), using Probabilistic Translation Dictionaries automatically created from parallel *corpora*. The process of bootstrapping an European Portuguese WordNet from the English, Spanish and Galician wordnets generated a total of 56,770 synsets and 97,058 variants.

TeP 2.0[24] (Maziero *et al*. 2008) is a vocabulary of Brazilian Portuguese, also developed from the Princeton wordnet framework, where words are grouped together in sets, establishing a synonymy relation (and in some cases antonymy), and providing examples for (some of) the queried words. The four major POS (noun, verb, adjective and adverb) are represented in TeP, which is freely available, and can be downloaded and used for many applications.

[18] There were in total 220 words associated to *pássaro* 'bird'. Due to this high number, only some few entries are shown here.
[19] The only information available is that the word belongs to the same *synset*. For consistency, we treated as a synonym.
[20] http://ontopt.dei.uc.pt/
[21] Namely the *Dicionário PRO da Língua Portuguesa 2005*, da Porto Editora, through the PAPEL project, the *Dicionário Aberto* (dicionario-aberto.net), Wikcionário.PT (https://pt.wiktionary.org), TeP 2.0 (http://www.nilc.icmc.usp.br/tep2/), the *OpenWordNet-PT* (https://github.com/own-pt/openWordnet-PT) and the *OpenThesaurus.PT* (http://paginas.fe.up.pt/~arocha/AED1/0607/trabalhos/thesaurus.txt).
[22] http://mwnpt.di.fc.ul.pt/
[23] http://wordnet.pt/
[24] http://www.nilc.icmc.usp.br/tep2/index.htm

Table 3: Semantically related entries for the keyword *amigo* 'friend'.
Proverb: *Quem te avisa, teu amigo é* 'He who warns you is your friend'

Language Resource	***amigo*** 'friend' **[semantic relation]: word(lemma)+**
WordNet.PT	[quasi-antonym of (noun)]: *inimigo* [co-related with (noun)]: *amizade* [hyponym (type of) (noun)]: *pessoa* [hypernym (supertype of) (noun)]: *amigalhaço*, *camarada*, *confidente* [quasi-antonym of (adjective)]: *inimigo*
PAPEL	[synonym_N_of]: *afeiçoado*, *aliado*, *amante*, *camarada*, *camba*, *chamuar*, *chapa*, *companheiro*, *partidário*, *simpatizante*, *xará* [synonym_ADJ_of]: *afeiçoado*, *aliado*, *unido* [hyperonym]: *amigalhote*, *antropófilo*, *copista*, *demófilo*, *filósofo*, *pendenciador*, *reinadio* [is a member of]: *pessoal* [antonym_ADJ_of]: *inimigo* [made_with]: *traição*
Onto.PT	[synonym (adjective)]: *afeiçoado*, *aliado*, *amante*, *amical*, *amigável*, *amigo*, *amistoso*, *apegado*, *apreciador*, *camarada*, *colaço*, *contubernal*, *ligado*, *pegado*, *unido* [synonym (noun)]: *achegado*, *acompanhante*, *aderente*, *afeiçoado*, *aliado amante*, *amásio*, *amizade*, *barregão*, *camarada*, *camba*, *capeba*, *chamuar*, *chapa*, *colega*, *companheiro*, *companhom*, *compincha*, *concubino*, *confrade*, *conhecido*, *contubernal*, *dedicado*, *inclinado*, *irmão*, *malungo*, *miga*, *mirmidão*, *partidário*, *simpatizante*, *xará*
MWN.PT	[hyponym of] *criatura*, *indivíduo*, *pessoa*, *ser humano* [is a kind of] *alter ego*, *camarada*, *colega de apartamento*, *colega de escol*a, *colega de quarto*, *compadre*, *companheiro*, *compincha*, *cupincha* (BR), *confidente*, *irmão*
PULO	[synonym] *aliado*, *aficionado*, *amante*, *amizade*, *colega*, *companheiro*, *irmão*, *seguidor*, *sócio*
TEP 2.0	[synonym (adjective)]: *afeiçoado*, *aliado*, *amical*, *amigável*, *amistoso*, *apegado*, *camarada*, *colaço*, *contubernal*, *ligado*, *pegado*, *unido* [synonym (noun)]: *afeiçoado*, *amante*, *amásio*, *camarada*, *companheiro*

Since most of the existing gaming exercises with proverbs feature several semantic relations between the distractors and the target word, we tried to assess the usefulness of some of these language resources in view of the automatic generation of didactic exercises, since they explicitly encode semantic relations between words. For two very common proverbs, appearing simultaneously in the 4 collections of proverbs we used to establish our *corpus* (section §4), we retrieved the words (potential distractors) these resources could produce when queried for entries that are in some semantic relation with the proverb keywords (the targets). The results are shown in Tables 2 and 3.

One of the issues that has to be addressed in the automatic generation of distractors for this type of exercise in the gender-number agreement between, on one hand, the target word (*v.g.*, *pássaro* 'bird': masculine-singular) and the surrounding words in the proverbs (*v.g.*, *um* 'a/one': masculine-singular, *dois* 'two': masculine-plural); and, on the other hand, the gender-number of the distractors (*v.g.*, *ave* 'bird': feminine-singular). Therefore, when retrieving the distractors, either these gender-number features act as filters, selecting only those words that match the target word values; or some the surrounding words of the proverb may have to be adjusted (or the distractor inflected appropriately) to keep the sentence grammatically correct. For example, the two proverbs would need to be adapted to produce the following, grammatically correct strings (distractors shown [inside square brackets] and modified word in **bold**):

*Mais vale **uma** [ave] na mão do que **duas** a voar*
'It is better to have a/one bird in the hand than two flying'

This latter option requires, however, the sentence be previously parsed in order to determine which words feature the gender-number agreement with the target (basically determiners and modifiers, or the verb if the target word is the head of the subject). This is made more complex when

non-local constraints are in place, as in this case, where *anaphora resolution* (Mitkov 2002, Marques 2014) is required: the numeral *dois* is an anaphor introducing a repeated (and, therefore, zeroed) instance of the target noun[25]:

*Mais vale um **pássaro** na mão do que dois* (**pássaros**) *a voar*

In this case, only one masculine noun is present in the sentence, that can be the antecedent of the zeroed word. However, in the case of the distractor *ave* 'bird', when producing the morphosyntactic adjustments for the gender-number agreement, two possible antecedents are available, the distractor proper and the word *mão* 'hand', also a feminine noun.

*Mais vale uma **ave** na mão do que duas* (**aves**) *a voar*

On top of it all, this inflection information is not available in lexical resources, since their purpose is different. Therefore, it is necessary to associate that information to the data retrieved from the ontologies, both for parsing the proverbs and to generate the adequate inflected forms for the morphosyntactic adjustments. Since there are already several computational lexicons of Portuguese (for example, Bick, 2000; Ranchhod *et al*., 1999, Vicente, 2013, among others), this should not constitute an insurmountable subtask, though it involves some additional processing.
On the other hand, for the parsing stage, the natural language processing required is much more complex and only some few systems exist, for Portuguese, that could produce such data. Among others, one could cite PALAVRAS (Bick, 2000), LX-Parser (Silva *et al*. 2010) and STRING (Mamede *et al.*, 2012).

Another issue comes from the fact that some of the items retrieved from the ontologies under the relation of synonymy are not exactly so, probably because some of these resources were automatically derived from existing dictionaries (for human users). For example, the following words were indicated as synonyms of *amigo* 'friend': *conhecido* 'acquaintance', *irmão* 'brother', *sócio* 'partner', *aficionado* '*id*.', but only those that can appear a predicative (post-copula) context with a human determinative complement (*X be N of Y*) are adequate to replace the target word. Thus, *aficionado* 'id.' does not have such properties and, therefore, it can not enter the proverb's structure (**Quem te avisa teu* [*aficionado*] *é*). On the other hand, for antonyms, one finds *inimigo* 'enemy'. This produces a semantic drift that may engender acceptability issues is the resulting distractors.

Naturally, some resources have limited usefulness since they do not feature the target words, or do not include any items from the same POS as the target. For example, TeP 2.0 has no entry for *pássaro* 'bird', a very common word, and MWN.PT only contains nouns, so it is useless for replacing verbs.

4. *CORPUS*

Several (printed!) collections of proverbs are already available in the literature, but to the best of our knowledge, only one digital resource exist, with about 2,300 proverbs and (some) variants[26]. To fill in this important gap in available linguistic resources, a *corpus* with +114,000 proverbs and their respective variants has been built (Reis, *in preparation*), digitized from four printed collections of proverbs (Costa, 1999; Machado, 1996; Moreira, 1996; Parente, 2005) and semi-automatically corrected. Each proverb has been identified unambiguously by a conventional code, allowing the user to recover the source it came from.

One of the first tasks to make this new *corpus* useful consisted in determining which proverbs appeared in more than one of the source collections. This aims at identifying the most usual proverbs, since different authors would have collected them more than once and independently. However, achieving that goal is not trivial, since proverbs present a wide variation both orthographic and in the

[25] In fact, this is a special type of anaphora, the so-called *lexical anaphor*, since the zeroed *anaphor* refers to the *lexical item* (the noun) in the antecedent *and not* to the same extra-linguistic entity that word refers to, as in a typical anaphoric relation.
[26] http://natura.di.uminho.pt/jjbin/dac

use of punctuation. Other proverbs are clearly variants, only differing slightly, in the use of a determiner, a preposition, or other grammatical words. Still, allowing for this difficulty, 44,272 (approximately 39%) of the proverbs appear more than once in the *corpus*, while 70,139 (approximately 61%) appear only once. Due to those variation factors (and remaining scanning/optical character recognition errors), it is likely that the number of repeated proverbs might increase.

Table 4: Distribution of the proverbs' keys by frequency classes (bins) in the *corpus*

Frq	Count	%	Frq	Count	%	Frq	Count	%
1	20,453	0,39	8	108	0,00	15	0	0,00
2	12,261	0,24	9	53	0,00	16	0	0,00
3	10,805	0,21	10	22	0,00	17	0	0,00
4	6,184	0,12	11	9	0,00	18	0	0,00
5	1,132	0,02	12	7	0,00	19	1	0,00
6	470	0,01	13	6	0,00	20	0	0,00
7	273	0,01	14	1	0,00	total	51,785	1,00

Since our ultimate goal is to group together variants of proverbs in order to be able to take that set of forms as a single *paremiologic unit*; and, as we intend to do so by associating to each proverb a set of distinctive key-elements, in order to be able to identify proverbs univocally in texts; we proceed with the construction of a **key** to each instance of the *corpus*, by automatically performing a set of operations on the strings of words that make up that large list. These operations included: (a) convert all text to lowercase; (b) remove all punctuation marks; (c) "cleaning" all repeated white spaces and incorrectly split words (due to the OCR); and (d) removing all *stopwords*[27], after establishing a list of word forms, adequate to the task at hand, from extant lists of stopwords for Portuguese available in the internet. The resulting list of keys contains 51,785 different strings, whose distribution is presented in Table 4.

In spite of its approximate nature, this method allowed, nevertheless, to group together several variants of the same paremiological units, simply by associating them to a key formed of their lexical elements (content words).

5. AUTOMATIC RECOGNITION OF PROVERBS IN TEXTS

To produce certain exercises, it is necessary to be able to retrieve automatically from *corpora* or from the web the texts where the proverbs naturally occur. Recently, Rassi *et al*. (2014) presented a formal (syntactic) classification, based on a *corpus* of 3,500 proverbs and their variants, organized in 595 types (or base forms) and collected from several dictionaries of Brazilian Portuguese proverbs. The authors presented a method to automatically produce finite-state machines using the Unitex linguistic development platform (Paumier, 2003, 2016) that can be used to retrieve candidate proverbs from texts. The method was tested in the PLN-Br *corpus* (Bruckschen *et al.* 2008), with 29 M tokens of journalistic texts, taken from the on-line edition of the Folha de São Paulo newspaper (1994-2005). A precision from 60% to 73% was reported, depending mostly on the formal class of the proverb and the degree of coverage of the variants represented in the list of proverbs. However, the number of proverbs found in this *corpus*, considering the total number of proverbs in the list, is quite small, which probably stems from the journalistic nature of the *corpus*.

Considering the data and the results obtained from this work, namely its reduced coverage of the language stock of proverbs, the issues found in the automatic generation of the finite-state automata, and their impact in the overall precision of the method, another approach has been envisaged. Based on the frequency of the proverbs' keys (produced as described in §), we intend to produce, manually, the finite-state automata (FSA) containing, for each paremiological unit the strings of keywords that uniquely identify it.

To better illustrate the method, let us consider the proverb *Deus escreve direito por linhas tortas* 'God writes straight with crooked lines', which is represented in the database by the following entries (the codes on the left represent the source of the variant), in alphabetic order.

[27] https://en.wikipedia.org/wiki/Stop_words; Witten *et al*. (2011).

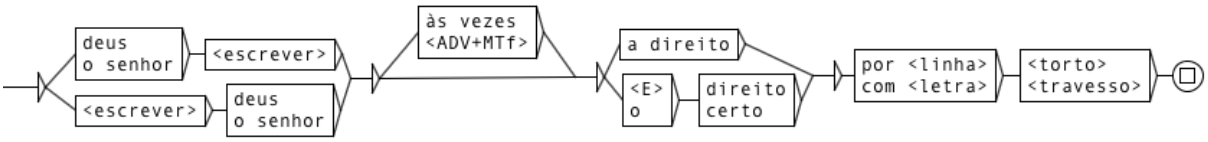
LPP_JRMC10132	*Deus escreve por linhas tortas.*
LP_SP11372	*Deus escreve direito (certo) por linhas tortas (travessas).*
GLP_JPM07945	*Deus escreve direito (certo) por linhas tortas.*
PP_AM03569	*Deus escreve direito por linhas tortas.*
GLP_JPM09795	*Escreve Deus às vezes o direito com letras tortas.*
PP_AM04397	*Escreve Deus às vezes o direito por linhas tortas.*
LP_SP14248	*Escreve Deus às vezes o direito por linhas tortas.*
LPP_JRMC10139	*Escreve Deus às vezes o certo por linhas tortas.*

The following variation can be found in these entries can be summarized as follows:

- Subject inversion: *Deus escreve/escreve* Deus 'God writes/writes God';
- Lexical variation: *Deus/o Senhor* 'God/the Lord'[28], *direito/certo* 'straight', *linhas/letras* 'lines/letters', *tortas/travessas* 'crooked'; and the equivalence of the above with the compound manner adverb *a direito* 'straightly'[29];
- Nominal use of adjectives *direto/certo* 'straight', indicated by the article *o* 'the';
- (facultative) insertion of frequency adverb *às vezes* 'sometimes'.

Considering the variation factors above, the following FSA can be built, using the Unitex (Paumier, 2003, 2016) linguistic development platform (Fig. 4)[30]:

Fig. 4: Graph with a finite-state automata representing the formal variation found in proverb *Deus escreve direito por linhas tortas* 'God writes straight with crooked lines'.



By systematically exploring the graph paths, one can obtain around 120 different strings. A simple query in a search engine of some of these variants is sufficient to retrieve hundreds of instances of this proverb from the web.

Though this is a painstaking process, particularly when there are several, very similar, keys associated with the same proverb, it is possible, in this way, to group together a large number of variants of a paremiological unit, using the large database of collected proverbs described above as source of variation, allowing some degree of free morphologic e syntactic variation (lemmatization, lexical masks, facultative insertions, punctuation), thus aiming at achieving a better performance both in precision and recall in the retrieval of proverbs from texts.

This description work is still at its beginning. Once the coverage of this FSA library is sufficiently large, we expect to use it to retrieve proverb-candidates from texts from real texts, in *corpora* of different genres and text type, as well as from the web. This will allow us to determine the relative frequency of proverbs' use, drawn from large textual *corpora*, which is an indispensable step towards the construction of didactic applications.

[28] This variant is not in the *corpus* of collected proverbs, but we found 2 instances in the web (using Google, with exact match and only under the top domain .pt).
[29] *Idem*: 6 instances of this variant were also found in the web.
[30] In the graph, the lemma of a word is given inside '<' and '>' (*v.g. escrever*, *linha*, *letra*, *torto* and *travesso*).This allows the system to match any inflected form associated with these lemmas. Notice also that *deus* and *senhor* are written in lowercase to allow case variation, should it appear in texts. The lexical mask `<ADV+MTf>` represents any adverb of the syntactic-semantic class of temporal (frequency) adverb, both simple and compound (Molinier and Levrier, 2000) Such lexical resources, with the syntactic-semantic information associated with adverbs, have already been built also for Portuguese (Palma 2009, Mamede *et al.* 2012).

We expect, in this way, to be able to collect a *corpus* of proverbs in their context of use (full texts), thus paving the way to a study on the linguistic devices that introduce proverbs into a discourse, and to model the complex discursive relations they hold with the text around it, linked with the communicative, situational-pragmatic factors surrounding those utterances. Notice that, besides their cultural aspects, this communicative competence is one of the key interests in learning proverbs, both as a native and (which more difficult) as a foreign language.

6. DIDACTIC GAMES: SOME EXPERIMENTS

In this section, we sketch some of the didactic games involving proverbs that can be produced using NLP tools and available resources for Portuguese. Adopting the general framework of REAP.PT, illustrated by the works of (Correia 2010; Pellegrini *et al*. 2012; Correia *et al*. 2012), and briefly outlined above (§), our purpose, here, is just to create a roadmap, defining the games' goals, and highlighting the complexity of the NLP tasks involved in automatically generating such exercises, eventually drawing the examples from real texts, whenever appropriate.

Most games already available in the web have been manually produced by specialists (language teachers) in very small and closed batches. These exercises are mostly *multiple-choice* (or *cloze*) *questions*. Each exercise can be seen as a HIT (Human Intelligence Task)[31] where the student, using a complex interplay of his language competence and world knowledge, has to choose the appropriate word or expression to fill in a blank in the proverb that provided as *prompt*. Besides the correct answer, the student is presented with 3 alternative answers -- called *distractors* (or foils), to choose from. The selection of the foils is a complex issue, deriving from the objectives set for a specific exercise. In the case of language learning, these objectives are related to the training of linguistic competences, mostly of semantic nature, such as the semantic relations holding between the target word/expression and the set of foils within the proverb.

Since the ultimate goal of this work to to automatically generate the exercises, NLP tools and language resources are necessary to produce the exercises, in such a way that a large number of HIT be built, stored in a quiz database and made available to the student. We will ignore, in this paper, all technical aspects concerning the database construction and management, as well as the web interface and the procedures for correction, scoring, and keeping track of the students answers for future reference, and will focus solely on the linguistic aspects that building such games raises and the challenges they pose to a NLP, automatic generation approach. In Section 3, we presented some of the linguistic issues in choosing the appropriate distractors for a given target word, based on the semantic relations it may hold with other lexical items, using the information encoded in existing ontologies already built for Portuguese. We resume that exercise here, adding some further examples.

Consider the well-known and already mentioned proverb, with the target word [*pássaro*]:

Mais vale um [***pássaro***] *na mão que dois a voar*
'Better a bird in the hand than two flying'

Multiple, and complex, linguistic competences are at play and not all are necessarily under the scope of a HIT based on this sentence:

- the ability to **parse** the syntactic structure of a complex sentence with a subordinate comparative subclause (*mais C1 do que C2*);
- Parsing the main clause *C1* also involves: (a) the reconstruction of a reduced infinitive *Vinf w* = [*ter um pássaro na mão*] as the subject of *vale mais*; (b) the reconstruction of an indefinite (hence zeroed) generic subject of *ter* (*alguém/toda a gente ter um passaro na mão* 'for someone/anybody/everybody to have a bird in the hand'), this type of subject and its zeroing being responsible for the 'generic', that is, 'proverbial' value of the proverb;

[31] https://en.wikipedia.org/wiki/Amazon_Mechanical_Turk#Artistic_and_educational_research

- while the parsing of the subordinate *C2* involves: (a) the (lexical) anaphora resolution (*dois pássaros*), explained above; and (b) reductions of repeated material under the comparative conjunction *do que* 'than'; an adequate analysis of the subordinate infinitive (*a Vinf = a voar*),
- as well as the knowledge of the **distributional constraints** on the subject of the verb *voar* in *C2* and of the complex clause underlying the relation between *pássaro* and *mão* in the first clause, involving locative preposition *em* and the particular grammatical value of *ter*.

It is on the latter, more semantically-oriented, aspect that the type of HITs here suggested hinge, though all other factors be involved as well. The following distractors can be automatically selected using the semantic relations encoded in the ontologies and replacing the the target word by:

- a synonym (*ave*) or a quasi-synonym (*piu-piu*)
- a hypernym (º*animal*, *vertebrado*)
- a diminutive of any of the words, including the target (*passarinho*, *ºavezinha*)
- a hiponym (*andorinha, arara, ...*)

All the examples above, except those marked with 'º', can be directly retrieved from the databases.

The specific selection of the noun *pássaro* in the proverb and using its synonym *ave*[32] as a distractor may not be entirely adequate, as 5 instances of this proverb with *ave* were found with Google exact match search (restricted to the top domain `.pt`), against 120 (different) instances of the variant with *pássaro*. A check for the full string (with all the selected foil-candidates, such as *ave*) in the web would prevent an intelligent system to produce incorrect foils (an alternative correct answer). Notice the issue does not rise with *piu-piu* and all other remaining options (no matches were found). Thus, such web search filtering mechanism should be of general use in any automatic exercise generation system. The choice of hypernyms conflicts with the somewhat constraint distribution found for the subject of *voar*. Besides, usually this strategy is not very productive, not only because of the limited number of lexical items functioning as hypernyms in the ontology hierarchy. The use of diminutives focuses not only on a morphologic competence, but also on the narrow distribution (or *frozenness*) of proverbs in general. It may be a good strategy for selecting foils, as diminutives (especially *-inho* and *-ito*) are a productive lexical device in Portuguese, which constitutes a further reason to be trained and learnt. Again, the filtering based on a web search would allow to discard *passarinho* (13 instances in Google) against *avezinha* (no exact match found). Finally, the list of hyponyms being very large would enable a substantial set of alternative foils.

Combining the different strategies in the same HIT could enable a system to produce a large number of HITs, all based on the same proverb. Irrespective of the computational complexity involved, all the necessary language resources and NLP tools exist and are adequate.

Another venue of research could result from exploring the fact that many proverbs are based on opposing/contrasting meaning, that is, words in an *antonymy* relation. However, in order to be able to explore such feature, one must first detected such proverbs having to antonym words.

To do so, we used the 388 word pairs of adjectival antonyms (*e.g.* the adjectives *direito*/*torto* 'right,straight'/'crooked,twisted') encoded in PAPEL (Oliveira *et al*. 2008), and queried all the entries in the proverbs database containing such pairs, irrespective of their order: 556 instances of proverbs (and variants) were found in the database, which correspond to 330 different strings. Below are ten of the first proverbs (sorted alphabetically) of that list, involving the pairs of adjectives *fácil/difícil* 'easy/difficult', *certo/incerto* 'certain'/'uncertain', f*eliz/infeliz* 'happy'/'unhappy', *amigo/inimigo* 'friend'/'enemy' and *possível/impossível* 'possible'/'impossible'

[32] We do not discuss here the issue of *pássaro* being a synonym, a hypernym or a hyponym of *ave*, which is a matter of linguistic adequacy of the semantic description, and pragmatically assume as correct any information in the language resources as given.

(1) *A crítica é **fácil**, a arte **difícil**.* 'Criticism is easy, art (craft) is difficult'
(2) *A crítica é **fácil**... a arte (é) **difícil**.* 'Criticism is easy, art (craft) is difficult'
(3) *A hora é **incerta** mas a morte é **certa**.* 'The hour is uncertain but death is certain'
(4) *A hora é **incerta**, mas a morte é **certa**.*'The hour is uncertain but death is certain'
(5) *A inconstância da fortuna assusta os **felizes** e anima os **infelizes**.* 'The fickleness of fortune scares the happy and animates the unhappy'
(6) *A lisonja faz **amigos** e a verdade, **inimigos**.* 'Flattery makes friends and truth enemies'
(7) *A marido **ausente**, amigo **presente**.* 'To absent husband, present friend (=lover)'
(8) *A melhor maneira de nos desfazermos de um **inimigo** é fazer dele um **amigo**.* 'The best way to discard an enemy is make a friend out of him'
(9) *A morte é **certa**, a hora **incerta**.* 'Death is certain, the hour uncertain'
(10) *A quem busca o **impossível**, justo é que até o **possível** se lhe negue.* 'To whom that seeks the impossible, it is fair that even the possible be denied to him'

Notice, first, that some of these instances correspond just to graphic (punctuation) variants of the same paremiological unit, namely (1) and (2) or (3) and (4); secondly, the reversal of the order of the antonymic word pair yields certain variants (9) where the two parts of the proverbs have just been permutated (3)-(4).

Let us consider again the proverb *Deus escreve direito por linhas tortas* 'God writes straight with crooked lines' and try to define possible HITs for this type of antonymy-based proverb. A very simple exercise would be to swap the adjectives:

Deus escreve torto por linhas direitas 'God writes crookedly by straight lines'

The purpose of such foil could be, for example, to identify ill-formed proverbs among correct variants. However, proverbs involving antonymy are very prone to creative reuse, which explains why 12 instances of such form have been found in the web, including a very well-known dramatic text[33]:

> Manuel: *Não é de espantar. Deus escreve torto por linhas direitas. Não é assim que se devia dizer?*
> 'Manuel: That does not surprise anyone. Gog writes crookedly by straight lines. It is not like this how one should say it?' (Luís de Sttau Monteiro, *Felizmente há luar!*)

It is also noteworthy that, if this permutation were performed, many of the proverbs of the list above, they would still look like proverbs, though the interpretation of some of them would become awkward:

*A crítica é **difícil**, a arte **fácil**.* 'Criticism is difficult, art (craft) is easy
*A hora é **certa** mas a morte é **incerta**.* 'The hour is certain but death is uncertain'
*A inconstância da fortuna assusta os **infelizes** e anima os **felizes**.* 'The fickleness of fortune scares the unhappy and animates the happy'
*A lisonja faz **inimigos** e a verdade, **amigos**.* 'Flattery makes enemies and truth friends'
*A marido **presente**, amigo **ausente**.* 'To present husband, absent friend (=lover)'
*A melhor maneira de nos desfazermos de um **amigo** é fazer dele um **inimigo**.* 'The best way to discard an enemy is make a friend out of him'
*A quem busca o **possível**, justo é que até o **impossível** se lhe negue.* 'To whom that seeks the possible, it is fair that even the impossible be denied to him'

[33] http://storamjoao.blogspot.pt/2008/11/anlise-de-felizmente-h-luar_29.html

On another perspective, one could also use the list of synonyms associated to each key adjective:

- ***direito*** (16 synonyms: *aprumado*, *desempenado*, *directo*, *erecto*, *especado*, *franco*, *imparcial*, *íntegro*, *justo*, *leal*, *liso*, *plano*, *rectiforme*, *recto*, *sincero*, and *vertical*)
- ***certo*** (19 synonyms: *ajustado*, *cabal*, *certeiro*, *claro*, *compassado*, *convencido*, *convicto*, *correcto*, *exacto*, *fixado*, *incontingente*, *indubitável*, *infalível*, *legal*, *positivo*, *preciso*, *seguro*, *tranquilo*, and *verdadeiro*)
- ***torto*** (16 synonyms: *desleal*, *embriagado*, *enviesado*, *errado*, *esconso*, *esguelhado*, *estrábico*, *inclinado*, *injusto*, *oblíquo*, *torcido*, *torso*, *tortuoso*, *troncho*, *vasqueiro*, *vesgo*)
- **travesso** (32 synonyms: *amarotado*, *atratantado* <sic>, *atravessado*, *colateral*, *desenvolto*, *diabólico*, *diabril*, *endemoninhado*, *endiabrado*, *estouvado*, *gabiru*, *gaiato*, *garoto*, *inquieto*, *irrequieto*, *judio*, *levadinho*, *levado*, *magano*, *maldoso*, *maroto*, *mexelhão*, *oblíquo*, *rabeador*, *rabino*, *remexido*, *roberto*, *safado*, *sapeca*, *traquinas*, *turbulento*, *vivo*)

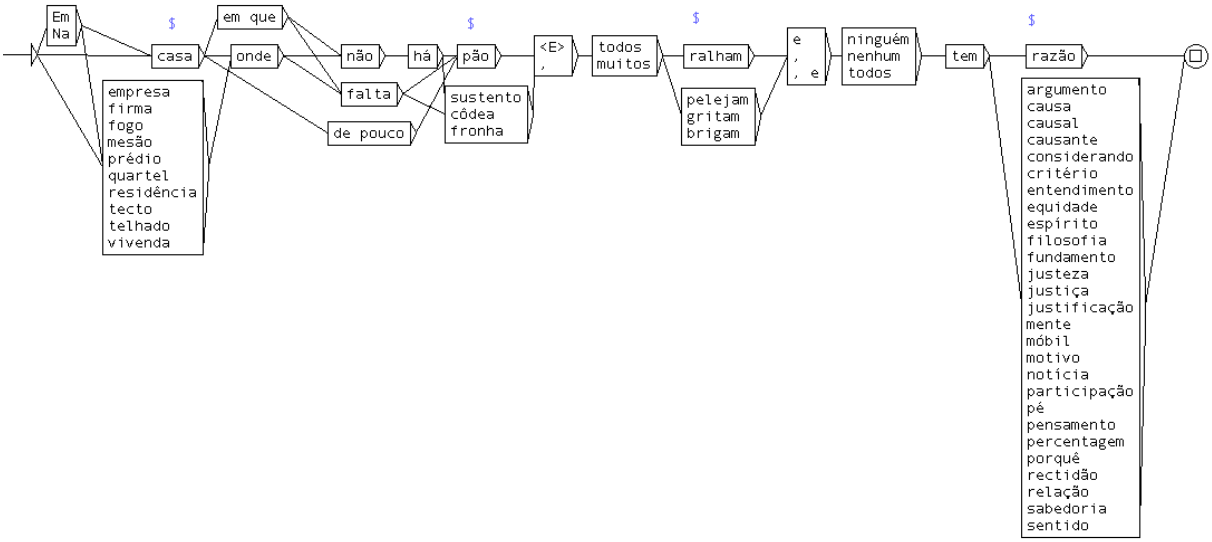
To create new distractors, it suffices to replace the key adjectives, one at time, keeping the others constant, by their synonyms This relatively simple permutation would yield 3,260 distractors (though not all were correct), that would enable the student to practice his/her semantic competence about the antonymy relations each adjective may (or may not) hold with the other adjectives in the proverb.

This, however, may be different depending on the language variety one considers. For example, one can imagine that in Brazilian Portuguese (but not in standard European Portuguese!) the equivalence of the (manner) adverbial use of the adjectives *direito* = *reto* and eventually *direto* would lead to consider as natural the variants (we found instances of both int the web):

Deus escreve reto/direto por linhas tortas ‘God writes straight by crooked lines’

Another type of exercise could involve the simultaneous variation of more that one word at the same time, which would increase the level of difficulty of the HIT. For example, consider the proverb *Em casa onde não há pão, todos ralham e ninguém tem razão* ‘In a house without bread, everyone argues and no one is right’. Defining as target words the nouns *casa* ‘house’, *pão* ‘bread’ and *razão* ‘reason’ and the verb ralhar ‘argue/discuss/quarrel’, and using their equivalent words, marked as synonyms in PAPEL, would yield the combinations (several thousands) represented in the graph of Fig. 4.

Figure 4: Multiple word variation in a single proverb: synonyms associated to four target words in the proverb *Em* [*casa*] *onde não há* [*pão*], *todos* [*ralham*] *e ninguém tem* [*razão*] ‘In a house without bread, everyone argues and no one is right’.



Notice that for each noun, the semantic relations between the target word and the so-called equivalent word are not always the same, depending on the word sense. By exploring this sense variation, the student may improve his/her lexical competence, extending the vocabulary not just in the number of new words, but also in the number of senses he/she can associate to already known words. For example, for *casa* 'house', the professional sense of *empresa* and *firma* 'company' is not to be confound to the sense of type of habitation building like *vivenda* 'villa' or *prédio* 'apartment building', nor with the figurative use of *tecto* and *telhado* 'roof', nor even with the almost technical value associated to *fogo* (literally: 'fire') 'habitation'. Some of these "equivalent" words are just plainly bizarre, probably an error resulting from the automatic methods used to build the ontology and the lack of human verification (e.g. *pão* 'bread' and *fronha* 'sleeping pillow cover'). Notice the figurative use of *pão* 'bread' in the equivalence with *sustento* 'sustenance/livelihood'. The case of *razão* 'reason' is more interesting, since this use corresponds to a support verb (or light verb) construction, *ter razão* 'to be right', and the lack of equivalence results from the simple word-by-word equivalence (and the many senses the simple noun may present), without considering this multiword lexical unit. Probably, targeting this noun would result in uninterpretable HITs. However, very few NLP systems, nowadays, can parse support verb constructions adequately (Rassi *et al*. 2014b).

Looking beyond antonymy and synonymy, other semantic relations can be explored in the automatic generation of foils for a proverb, namely hiperonymy and hiponymy, or *whole-part* relations, i.e. holonymy and meronymy. An example of the later might be construed for the proverb:

> *Mais vale um pássaro na mão do que* [*dois* (*pássaros*)] *a voar*
> 'It is better to have a/one bird in the hand than two flying',

where the numeral *dois* (pássaros) 'two (birds)' in the headless noun phrase could be replaced by *um bando* (*de pássaros*) 'a flock (of birds)' and then by different types of collective nouns, to explore the lexical knowledge about the distributional constraints on the use of collective nouns designating groups of animals (holonymy):

> *Mais vale um pássaro na mão do que* ***um bando/cardume/cacho/chorrilho*** *a voar*
> 'It is better to have a/one bird in the hand than a flock/school/bunch/ flying'

Naturally, we assume that the issue of anaphora resolution should has already been solved.

Finally another, rather challenging, exercise could involve discovering the appropriate proverb for a given context. However, instead of doing like some of the websites that explain the proverbs' meanings by presenting examples produced by their authors, one would use NLP tools to retrieve from the web and other sources real, spontaneous instances of proverbs in their context of use. This requires being able to finding proverbs in texts, especially the most frequent, and hence the better known, variants, and selecting those instances more prone to the generation of adequate distractors. For example, one of the proverbs already mentioned above has been found in a blog[34], with an extensive enough context to provide some of the clues required for guessing it:

> *Estudos feitos revelam que mais de 50% dos conflitos conjugais acontecem devido a problemas financeiros. Há um ditado popular que diz "*[*Em casa onde não há pão todos ralham e ninguém tem razão*]*", e será que não é verdade?*
> 'Studies show that more than 50% of marital conflicts happen because of financial problems. There is a popular saying that goes like "In the house where there is no bread all quarrel and no one is right," and is it not true?'

In the preceding text, several nouns can serve as clues for this proverb: *conflitos conjugais* 'marital conflicts' and *problemas financeiros* 'financial problems'. Other content words, like *estudos*

[34] http://reorganiza.pt/guia-orcamento-familiar-eficaz/

'studies', *verdade* 'truth' or even the verb *revelar* 'reveal', have little to do with the proverb's lexical content. Furthermore, the proverb is clearly identifiable by the devices used to introduce it, namely, the quotation marks and the formula H*á um ditado popular que diz* 'There is a popular saying that goes like' (this type of formula, if identified, could be ruled out from the set of clues,

Different strategies can be put in place. A very simple approach is to retrieve from the *corpus* of proverbs the entries that have two or more of the same words of the contextual clues. In our *corpus*, there is a instance of such situation (though it is not clearly a proverb), where the word *problema* appears repeated (no case was found with any of the word pairs):

> *Os problemas existem para se lhes dar soluções; quem não quiser defrontar os trabalhos de soluções deve evitar os problemas.*
> 'Problems are there to be given solutions; who does not want to face the work required for the solutions should avoid any problems.'

Alternatively, from an ontology like MultilingualWordNet.PT[35] one can retrieve the words that have a given semantic relation with the contextual clue (Table 5), much as already shown above. As one can see, some words may be missing from the ontology. It is then relatively simple to extract from the *corpus* the proverbs containing two or more words from the combinations of these elements. For example, one finds:

> *A desgraça do homem é falar fino, esmorecer, brigar com a mulher e ficar perto.*
> 'The man's misfortune is talking with a thin voice, fading/waning, fight with his wife and stay close.'

where the mention of a marital relation and the verb *brigar* 'quarrel' (but no mention of *problema* 'problem' nor *financeiro* 'financial') make this a good distractor. However, most of these word combinations may not exist in the *corpus*.
A more sophisticated approach would be to compute the word sense similarity (Clark *et al*., 2010) between the words in the proverb, on one hand, and those of the context, on the other hand, thus retrieving from the *corpus* the proverbs that are similar/dissimilar with the values obtained.

Table 5: Semantic relations in MWN.PT

Word	**[Semantic relation]**
estudo	[hyperonym (is a kind of)]: *composição musical*, *música*, *peça musical*
conflito	[is a word for]: *briga*, *combate*, *luta* [hyperonym (is a kind of)]: *estado*, *acto colectivo* (PE), *ato coletivo* (PB) [hyponym]: *briga*, *choque*, *combate*, *contenda*, *desacordo*, *desavença*, *desentendimento*, *desinteligência*, *dissenção*, *divergência*, *duelo*, *embate*, *enfrentamento*, *fricção*, *guerra fria*, *lide*, *luta*, *pancadaria*, *peleja*, *pugna*, *recontro*, *refrega*, *rixa*
conjugal	not found
financeiro	not found
problema	[is a word for]: *aperto*, *apuro*, *dificuldade*, *problema* [hyperonym (is a kind of)]: *acontecimento*, *sucedido* [hyponym]: *aflição*, *atrocidade*, *adversidade*, *desdita*, *desgraça*, *desventura*, *escândalo*, *indignidade*, *constrangimento*, *embaraço*, *inferno*, *interferência*, *massacre*, *ultraje*

[35] http://mwnpt.di.fc.ul.pt/ (we disregard the synset structure).

7. CONCLUSIONS AND FUTURE WORK

In this paper, we present some available NLP tools and resources that can be used to develop didactic exercises, automatically generated and corrected, to train linguistic competence students of PFL (and also of native speakers) and built around proverbs. We demonstrated that existing exercises require an inordinate amount of time and effort to be sufficiently apt to train those competences, so the automation of the process of creating them, particularly in a gaming environment, would help students and teacher alike, and improve the learning process. We have shown that these existing exercises heavily rely on semantic relations among words. Portuguese already has a number of automatically built or manually crafted language resources, as well several NLP systems that could be use as the basis for such project. The iCALL framework provided by the REAP.PT project could be adopted for such project. Naturally, this would require a close interaction and collaborative work involving both linguists and computer scientist. Part of the preliminary work is already under way: a large *corpus* of proverbs and variants has been collected and the construction of tools for retrieving paremiologic units (proverbs and their variants) from text is being done. In the near future, we expect to present some results from this process.

AKNOWLEDGMENTS

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

REFERENCES

Bick, E. (2000). *The Parsing System "PALAVRAS": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Dr.phil. Thesis. Aarhus University. Aarhus, Denmark: Aarhus University Press.

Bruckschen, M.; Muniz, F.; de Souza, J.G.C.; Fuchs, J. T.; Infante, K.; Muniz, M.; Gonçalves, P.N.; Vieira, R. and Aluísio, S. (2008). *Anotação linguística em XML do corpus PLN-BR*. Technical Report. São Carlos (SP).

Chacoto, L.M.V.G. (1994). *Estudo e Formalização das Propriedades Léxico-Sintácticas das Expressões Fixas Proverbiais*. Master thesis, Faculdade de Letras da Universidade de Lisboa.

Clark, A.; Fox, C. and Lappin, S. (eds.). (2010). *The Handbook of Computational Linguistics and Natural Language Processing.* Wiley-Blackwell.

Correia, R.P.S. (2010). *Automatic Question Generation for REAP.PT Tutoring System*, Master Thesis, Universidade Técnica de Lisboa, Instituto Superior Técnico.

Correia, R.; Baptista, J.; Mamede, N.; Trancoso, I. and Eskenazi, M. (2010). Automatic Generation of Cloze Question Distractors. *In: Proceedings of the Workshop on Second Language Studies: Acquisition, Learning, Education and Technology*, Tokyo, Japan.

Correia, R., Baptista, J., Eskenazi, M. and Mamede, N. (2012). Automatic Generation of Cloze Question Stems. *In Proceedings of the 10th International Conference on Computational Processing of the Portuguese Language*. LNAI/LNCS 6001: 168-178. Berlin/Heidelberg: Springer.

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

Council of Europe. (2011). *Common European Framework of Reference for Languages: Learning, Teaching*. Council of Europe.

Fellbaum, C. (1998, ed.) *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Dias-da-Silva, B.C. (2013). Modelagem linguístico-computacional de léxicos. In: LAPORTE, Éric; SMARSARO, Aucione; VALE, Oto A. (Orgs.). *Dialogar é preciso: linguística para processamento de línguas*. 1st edition, Vitória: PPGEL/UFES: 89-103.

Gross, M. (1996). *Grammaire Transformationnelle du Français. Syntaxe de l'Adverbe.* Paris: ASSTRIL.

González Rey, I. (2002). *La Phraséologie du Français*. Toulouse. Presses Universitaires du Mirail.

Hoshino, A. and Nakagawa, H. (2005). A Real-Time Multiple-Choice Question Generation for Language Testing: A Preliminary Study. *In Proceedings of the Second Workshop on Building Educational Applications Using NLP, Association for Computational Linguistics*: 17-20.

Machado, J.P. (1996). *O Grande Livro dos Provérbios*, 1st edition, Lisboa: Editorial Notícias.

Mamede, N.; Baptista, J.; Diniz, C. and Cabarrão, V. (2012). STRING: An Hybrid Statistical and Rule-Based Natural Language Processing Chain for Portuguese. In: Caseli, H.; Villavicencio, A.; Teixeira, A. & Perdigão, F. (eds.), *Computational Processing of the Portuguese Language, Proceedings of the 10th International Conference, PROPOR 2012 Demo Sessions*, vol. Demo Session.

Marques, J.S. (2013). *Anaphora Resolution*. Master thesis, Engenharia Informática e de Computadores, Instituto Superior Técnico, Universidade Técnica de Lisboa.

Marrafa, P. (2001) *WordNet do Português: uma base de dados de conhecimento linguístico*. Lisboa: Instituto Camões.

Maziero, E.G.; Pardo, T.; Di Felippo, A. and Dias-da-Silva, B.C. A Base de Dados Lexical e a Interface Web do TeP 2.0 - Thesaurus Eletrônico para o Português do Brasil. *In VI Workshop em Tecnologia da informação e da linguagem humana (TIL)* (Vila Velha, ES, Brasil, 28-29 Outubro 2008): 390-392.

Mel'čuk, I. and Polguère. (2007). *Lexique actif du français - L'apprentissage du vocabulaire fondé sur 20 000 dérivations sémantiques et collocations du français*. Bruxelles. De Boeck.

Mitkov, R. (2002). *Anaphora Resolution*, Longman, Studies in Language and Linguistics.

Molinier, C. and Levrier, F. (2000). *Grammaire des Adverbes: Description des Formes em 'ment'*. Droz. Genève

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

Oliveira, H. G.; Santos, D.; Gomes, P. and Seco, N. PAPEL: a dictionary-based lexical ontology for Portuguese. In António Teixeira, Vera Lúcia Strube de Lima, Luís Caldas de Oliveira & Paulo Quaresma (eds.), *Computational Processing of the Portuguese Language, 8th International Conference, Proceedings (PROPOR 2008)* Vol. 5190, (Aveiro, Portugal, 8-10 de Setembro de 2008), Springer Verlag: 31-40.

Oliveira, H.G. and Gomes, P. (2014). ECO and Onto.PT: A flexible approach for creating a Portuguese wordnet automatically. In *Language Resources and Evaluation* **48(2)**: 373-39. Springer.

Palma, C. (2009). *Expressões Fixas Adverbiais Descrição léxico-sintáctica e subsídios para um estudo contrastivo Português-Espanhol*. Master thesis. Universidade do Algarve.

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

Paumier, S. (2003). De la Reconnaissance de Formes Linguistiques a l'Analyse Syntaxique. Volume 2, Manuel d'Unitex. Ph.D. thesis, IGM, Université de Marne-la-Vallée.

Paumier, S. (2016). *Unitex 3.1 - User Manual*. Université de Paris-Est/Marne-la-Vallée - Institut Gaspard Monge, Noisy-Champs.

Pellegrini, T.; Ling, W.; Silva, A.; Correia, R.; Trancoso, I.; Baptista, J. and Mamede, N. (2012). Overview of Computer-assisted Language Learning for European Portuguese at L2F. *In Proceedings of the 4th International Conference on Computer Supported Education - CSEDU*: 538-543. Porto, Portugal.

Perkins, D. (1999). The many faces of constructivism. *Educational Leadership*, **5(3)**: 6-11.

Pianta, E.; Bentivogli, L. and Girardi, C. (2002). MultiWordNet: developing an aligned multilingual database. *In Proceedings of the 1st International WordNet Conference, January 21-25, 2002, Mysore, India*: 293-302.

Ranchhod, E.; Mota, C. and Baptista, J. (1999). A computational lexicon of portuguese for automatic text parsing. *In Proceedings of SIGLEX99: Standardizing Lexical Resources, 37th Annual Meeting of the ACL*: 74-80. College Park, Maryland, USA.

Ranchhod, E.M. (2003). O Lugar das Expressões 'Fixas' na Gramática do Português *in Ivo Castro & Inês Duarte (orgs.) Razões e Emoção*, vol II. Lisboa: Colibri: 239-254.

Rassi, A.; Baptista, J. and Vale, O. (2014). Automatic Detection of Proverbs and their Variants. Leibniz (Germany): Schloss Dagstuhl, Leibniz, Zentrum fur Informatik, Dagstuhl Publishing: *Proceedings of the Symposium on Languages, Applications and Technologies (SLATE'14*): 235-249.

Rassi, A.P.; Santos-Turati, C.; Baptista, J.; Mamede, N. and Vale, O. (2014b) The fuzzy boundaries of operator verb and support verb constructions with dar "give" and ter "have" in Brazilian Portuguese. *Workshop on Lexical and Grammatical Resources for Language Processing (LG-LP 2014), COLING 2014*, At Dublin, Ireland: 92-101.

Reis, S. (*in preparation*). *Constituição de um corpus de provérbios e variantes para representação da variação formal de unidades paremiológicas* (Technical Report). Faro: UALG.

Rocha, P. and Santos, D. (2000). CETEMPúblico: Um *corpus* de grandes dimensões de linguagem jornalística portuguesa. I*n: Nunes, M.G. et al., eds., V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR'2000)*, São Paulo: ICMC/USP: 131-140.

Silva, J.; Branco, A.; Castro, S. and Reis, R. (2010). Out-of-the-Box Robust Parsing of Portuguese. In *Proceedings of the 9th International Conference on the Computational Processing of Portuguese (PROPOR'10)*: 75-85.

Simões, A. and Gómez Guinovart, X. (2014). Bootstrapping a Portuguese wordnet from Galician, Spanish and English wordnets. *In Advances in Speech and Language Technologies for Iberian*

Languages, Proceedings of 2nd International Conference, IberSPEECH 2014, Las Palmas de Gran Canaria, Spain, volume 8854 of LNCS: 239-248. Springer.

Vicente, A. (2013). *LexMan: um Segmentador e Analisador Morfológico com Transdutores*. Master thesis. Instituto Superior Técnico, Universidade Técnica de Lisboa, Lisboa, Portugal.

Witten, I.H.; Frank, E. and Hall, M.A. (2011). *Data Mining Practical Machine Learning Tools and Techniques*. 3rd edition, USA: Morgan Kaufmann Publishers Inc.