Syntactic REAP.PT

Exercises on Word Formation

Pedro Henrique Santos Figueirinha

Thesis to obtain the Master of Science Degree in
Information Systems and Computer Engineering

Examination Committee

Chairperson:	Prof. Doutor Mário Rui Fonseca dos Santos Gomes
Supervisor:	Prof. Doutor Nuno João Neves Mamede
Supervisor:	Prof. Doutor Jorge Manuel Evangelista Baptista
Member of the Comittee:	Doutora Paula Cristina Quaresma da Fonseca Carvalho

October 2013

Acknowledgements

I would like to thank both my advisors, Professor Nuno Mamede and Professor Jorge Baptista, for allowing me to develop my ideas, while providing countless suggestions and guiding me throughout this whole process.

I would also like to express my gratitude to all the members of the L^2F, for providing me with the tools and assets necessary to complete this work, as well as with valuable discussions and clarifications. I would like to thank Cláudio Diniz in particular for his availability and for helping me with the integration of the new exercises in the already existing platform.

To my colleagues at Instituto Superior Técnico, I would like to give my deepest thanks for their continuous support over the last few years, and for making sure I stayed on the right track. I would also like to thank André Silveira for helping me with some of the graphical assets.

Last but not least, I would also like to thank my family and my friends. Without their support, everything would have been much more difficult.

Lisbon, October 15, 2012

Pedro Henrique Santos Figueirinha

To my parents.

Resumo

A introdução de computadores na área da educação, como uma ferramenta auxiliar, tem tido impacto na forma como os professores planeiam as suas aulas e interagem com os seus alunos O Ensino Inteligente de Línguas Assistido por Computador (ICALL) é o campo que aborda esta evolução no processo educativo e tem gerado um grande interesse na comunidade académica, nas últimas décadas. O REAP.PT é um sistema português originário desta área de pesquisa, cujo objectivo é o de ensinar português a portugueses e a estrangeiros. Este processo é feito com recurso a um conjunto de exercícios gerados automaticamente, que abordam vários aspectos, sintáticos e semânticos, da língua portuguesa, especificamente adaptados às necessidades dos utilizadores. O objectivo deste trabalho é expandir o sistema REAP.PT mais concretamente, o módulo de exercícios sintáticos, através da introdução de dois novos exercícios, que dizem respeito aos processos de formação de palavras. Tal como acontece com os exercícios previamente desenvolvidos, estas novas adições manterão o estilo e a interface visual típica da plataforma, assim como a funcionalidade que permite a geração automática de exercícios. O primeiro exercício aborda o processo de divisão de palavras em sílabas, processo este que embora possa parecer simples e fácil, requer que um conjunto específico de regras seja seguido à risca. O segundo exercício consiste na identificação de palavras que não são formadas por prefixação, embora aparentem ser à primeira vista. A existência de feedback personalizado, também é um ponto comum a ambos os exercícios. Os resultados positivos obtidos na etapa de avaliação, que consistiu em testes online, validam a abordagem seguida.

Abstract

The introduction of computers as an assisting tool in the field of education has affected the way teachers plan their classes and interact with their students. The Intelligent Computer Assisted Language Learning (ICALL) field addresses this evolution in the teaching process and has fomented great interest for the last few decades in the academic community. REAP.PT is a Portuguese system from this research area and aims to teach the Portuguese language to both foreign and native speakers. This is done by employing a series of automatically generated exercises focusing on several semantic and syntactic aspects of the language, specifically tailored to the users needs. The goal of this work is to expand the REAP.PT system, namely, its syntactic exercises, by introducing two new ones focusing on word formation. As is the case with the already developed exercises, these will not only maintain the REAP.PT visual style and interface, but they will also be automatically generated. The first exercise deals with the task of dividing words into syllables, a process that even tough looks simple and straight-forward at first, requires specific rules to be followed. The second exercise focuses on identifying words which are not formed by prefix addition, even tough they look the opposite. Personalized support and feedback is also presented for each of the exercises. The positive evaluation results, consisting of crowd-sourced tests, validate the presented approach.

Palavras Chave
Keywords

Palavras Chave

Ensino da Língua Assistido por Computador

Língua Portuguesa

Exercícios Sintácticos

Geração Automática de Exercícios

Keywords

Intelligent Computer Assisted Language Learning

Portuguese Language

Syntactic Exercises

Automatic Exercise Generation

Contents

List of Figures

List of Tables

Acronyms

AWL Academic Word List is a list of words which appear with high frequency in English-language academic texts.

CALL Computer-assisted Language Learning is "the search for and study of applications of the computer in language teaching and learning".

CMU Carnegie Mellon University is a private research university in Pittsburgh, Pennsylvania, United States.

FCT Fundação para a Ciência e a Tecnologia is an organization within the Portuguese Ministry of Education and Science with its purpose being the funding of scientific research activities.

ICALL Intelligent Computer-Assisted Language Learning is an interdisciplinary research field integrating insights from computational linguistics and artificial intelligence into computer-aided language learning.

INESC-ID Lisboa Instituto de Engenharia de Sistemas e Computadores: Investigação e Desenvolvimento is a non-profit organization devoted to research in the field of information and communication technologies.

L^2F Spoken Language Systems Laboratory is a research department at INESC-ID Lisboa.

LTI Language Technologies Institute is a division of the School of Computer Science at Carnegie Mellon University, in the area of language technologies.

NLP Natural Language Processing is an interdisciplinary research field of artificial intelligence and linguistics that studies the processing and manipulation of natural language.

P-AWL Portuguese Academic Word List is the corresponding Portuguese version of the English Academic Word List.

REAP REAder-specific lexical Practice for improved reading comprehension is a tutoring system developed at the Language Technologies Institute (LTI) of Carnegie Mellon University (CMU) to support the teaching of a language for either native or foreign speakers, through the activity of reading and focusing the students in learning vocabulary in context.

REAP.PT REAder-specific Practice PorTuguese is the Portuguese version of the REAP system.

STRING STatistical and Rule-based Natural lanGuage is an NLP processing chain for Portuguese developed a L^2F.

TAGARELA Teaching Aid for Grammatical Awareness, Recognition and Enhancement of Linguistic Abilities is an ICALL system for the Portuguese language.

WERTi Working with English Real Texts is an ICALL system for the English language.

YAH Yet Another Hyphenator is an hyphenator for the Portuguese language developed at L^2F.

1 Introduction

With the increasing role of technology integration and computer usage in our society and daily life, more particularly in the educational area, language teaching has become an increasingly attainable activity to anyone with the means to use a computer. The way the contents from this specific field can be easily accessed, has allowed the average user to gradually improve his/her proficiency in both native and foreign languages, given that the variety of available studying materials can cover a limitless number of shortfalls and gaps regardless, once again, of the users level of knowledge and skills.

From the need to conveniently adapt these resources to the specific needs of the users, the Intelligent Computer Assisted Language Learning (ICALL) research area was born. It is an academic field solely devoted to improving and simplifying this learning process.

One of the systems that emerged in this field is the REAP (REAder-specific Practice) project[1]. Initially developed at CMU (Carnegie Mellon University) by the LTI (Language Technologies Institute), the main goal of this project is to teach English by using a series of different exercises which focus on vocabulary, however, given the way these exercises are planned and implemented, the users will also be steadily improving their reading ability.

One aspect that sets REAP apart from the existing alternative systems, is the ability to automatically generate exercises on a variety of linguistic aspects.

With this in mind, we can describe the main goal of the REAP.PT[2] project as to port and adapt the original REAP features to the Portuguese language. Funded by the FCT (Fundação para a Ciência e Tecnologia), it has been continually improved and expanded over the years. The first module to be developed by Marujo (Marujo, 2009) was the lexical learning component, consisting of the integration of various textual and multimedia resources. After this, a component capable of generating questions automatically was built by Correia (Correia, 2010). Following these implementations, syntactic exercises were developed and integrated by Marques (Marques, 2011) and Freitas (Freitas, 2012), an interactive game using a three dimensional engine, was created by Silva (Silva, 2011) and finally, an additional exercise focusing on speech recognition using the same previously mentioned engine was also partly developed by Silva (Trancoso et al., 2012).

[1] `http://reap.cs.cmu.edu` (last visited in September 2013)
[2] `http://call.l2f.inesc-id.pt/reap.public/` (last visited in September 2013)

The latest improvement to the REAP.PT system, was a complete redesign of its interface, by Filipe (Filipe, 2013). This was done to give the system, and its games, a more homogeneous and attractive visual style.

1.1 Goals

The goal of the present work is to extend the available exercise database by introducing and implementing new exercises focusing on unexplored topics and aspects of the Portuguese language. More specifically, the main objective is to expand the work developed by Marques (Marques, 2011) and Freitas (Freitas, 2012) in the syntactic area. This extension strives to uphold the quality standards and features exhibited by the exercises already developed, namely, the correctness from a linguistic point of view and the automatic generation of exercises. The resulting output of this operation is particularly tailored to the user's preferences and their level of proficiency in the given language (Portuguese in this case).

As is it the case with the exercises already available in the REAP.PT system, the exercises developed in this project also maintain the visual style defined by Filipe (Filipe, 2013), they support several degrees of difficulty and they offer dynamic feedback consisting on what the user answers. The architecture for both exercises is detailed in Chapter 4.

1.2 Document Structure

The present thesis consists of 5 Chapters. Their contents are:

- Chapter 2 introduces the REAP.PT system, describing its architecture and the current available exercises.
- Other CALL and ICALL systems, including Portuguese ones, are presented in Chapter 3. Also in this chapter, examples of exercises focusing on the topics covered in this thesis are shown. These were extracted from Portuguese textbooks.
- The architecture of the syllabic divisor and both exercises is described in Chapter 4. Both exercises are detailed using real generated examples.
- Chapter 5 covers the evaluation process, including its set-up and results.
- Finally, in Chapter 6, the conclusions of this thesis are presented, as well as some suggestions for future work.

2 REAP.PT

In this chapter, the overall architecture of the REAP.PT system is described. Some of its components, such as the Daily REAP.PT, are explained in greater detailed, as well as the already developed exercises.

2.1 REAP.PT Architecture

Initially ported for Portuguese from the original English REAP system by Marujo (Marujo, 2009), the REAP.PT system has been expanded over the last few years by several researchers who worked continuously to improve the platform. The REAP.PT project consists of several components. Figure 2.1 provides an overall view of the current system architecture and the components of the system.

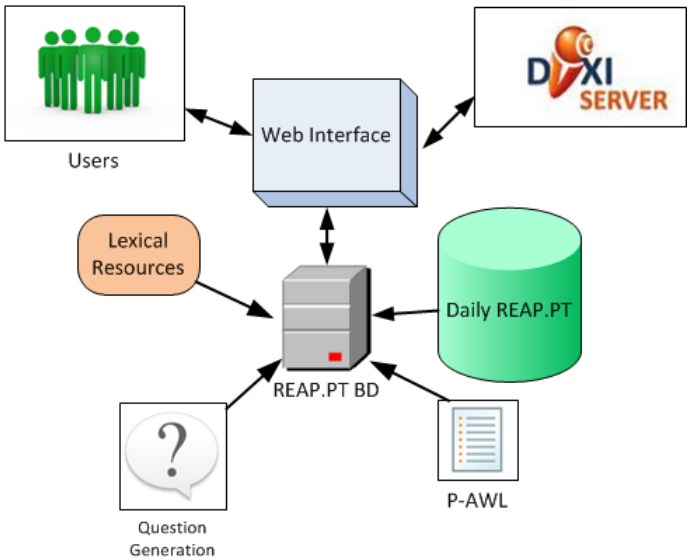

Figure 2.1: REAP.PT architecture.

The Web Interface module provides, to both teachers and students, the means to interact with the system. Furthermore, this component also allows the users to query the *Infopédia* on-line dictionary[1] for word definitions. Finally, the information exchange between the REAP.PT database and the DIXI Server (Paulo et al., 2008) – the listening comprehension module – is also handled by the Web Interface.

[1] `http://www.infopedia.pt` (last visited in September 2013)

Unique to REAP.PT and not present in the original English system, this module adds text-to-speech features to the system, enabling the users to playback any type of word or sentence from the given texts.

The Database component is actually composed of two separate databases. While the first one (also unique to REAP.PT) is used to keep a record of the current state of the system, storing information about the profiles of the users (their exercise history, interests, etc...), the second one is used to save the contents (dictionary, word relations, sample sentences and texts, etc...) extracted from various external lexical resources namely, PAPEL (Oliveira et al., 2008), MWN.PT[2] and TemaNet (Marrafa et al., 2006).

The Daily REAP.PT (Trancoso et al., 2012) component, which can be seen in greater detail in Figure 2.2, is a dataset consisting of texts gathered from Portuguese on-line newspapers and transcripts from television newscasts from the past seven days. To build this dataset, a web crawler is used to extract content from the world wide web. The Nutch web crawler (Moreira et al., 2007), was chosen for this purpose.

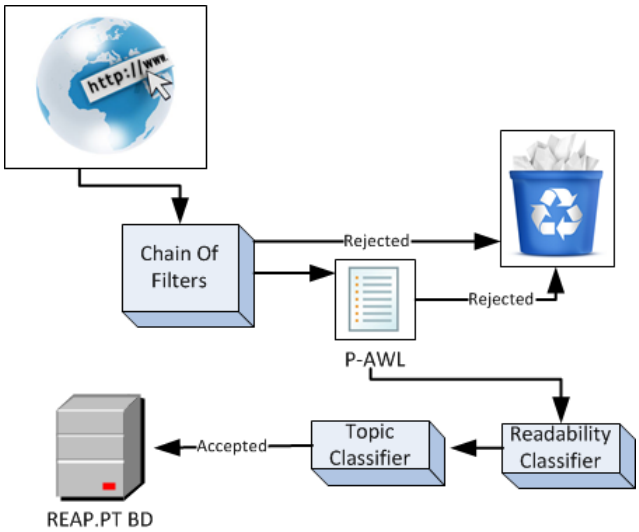

Figure 2.2: Daily REAP.PT component.

However, not all the texts extracted in this manner are useful for educational purposes, specifically, as teaching materials. They must be filtered to produce an output consisting of a smaller subset of texts, according to pre-specified selection criteria (Marujo, 2009). The filters that compose this chain are interconnected and the order by which they are applied is fixed. The first one limits the subset to texts larger than 300 words in length, automatically excluding the smaller ones. This operation also saves the length of each text in the REAP.PT database.

After this, texts containing profanity words are also filtered and lists of words are also excluded,

[2]Developed at the Natural Language and Speech Group of the Department of Informatics of the University of Lisbon, under the coordination of António Branco `http://mwnpt.di.fc.ul.pt/` (last visited in September 2013)

that is, texts that do not make sense and are seemingly just random assortments of words, not containing any valid sentences (e.g. sentences without verbs, sequences of identical parts of speech, etc...). The final filter to be applied has to do with the presence of words from the Portuguese Academic Word List (P-AWL) (Baptista et al.,) in the texts. The P-AWL is described by its authors as a "general purpose vocabulary, with current (but not colloquial) words, which has been designed for immediate application on a CALL web-based environment, currently devoted to improve students' reading practice and vocabulary acquisition". Any text that does not contain at least three words from this list is rejected.

After the filtering stage, accepted texts are classified based on their Readability, using the Collins-Thompson and Callan algorithm (Collins-Thompson & Callan, 2005), and on their Topic (Correia, 2010), making it possible for the users to select texts that fit in their interest areas.

The Question Generation module produces exercises that are presented to the students after they finish reading texts. These exercises consist of automatically generated multiple choice questions (Marujo, 2009; Correia, 2010), including definition questions, synonym/antonym questions, hyperonym/hyponym questions, simple cloze questions and open cloze questions, all based on the text read by the student.

The alternatives to the correct answer in each question (distractors) are also automatically generated and were developed and integrated in the system by Correia (Correia, 2010).

2.2 REAP.PT Vocabulary and Syntactic Exercises

REAP.PT offers several different exercises based on different aspects of the Portuguese language. These include vocabulary and syntactic exercises, with the later being composed of traditional text exercises developed by Marques (Marques, 2011) and Freitas (Freitas, 2012).

Using the system's interface, the user is able to provide data regarding their level of language proficiency and on their topics of interest, which is then stored in the database, together with the user's history of completed activities, which include reading sessions and exercise completion, along with parallel operations such as dictionary lookups. By combining the data from the different sources, texts are sorted and a specific collection of these is then presented to the user for both reading sessions and exercises.

The development of syntactic exercises took place after the implementation of manual cloze questions and the automatic generation of vocabulary and cloze questions, were done by Marujo (Marujo, 2009) and Correia (Correia, 2010) respectively. The later were implemented according to one of REAP's objectives, which is to provide the user with personalized exercises based on their level of proficiency.

Regarding vocabulary exercises, the focus is on learning a specific subset of words, properly

adapted to the user's level, which can be found in the Portuguese Academic Word List (P-AWL) (Baptista et al.,). These are also sometimes referred to as *focus words*.

Finally, with these aspects in mind, we can now describe each exercise.

2.2.1 Lexical Mahjong

The Lexical Mahjong exercise was developed by Marques (Marques, 2011) and like the original Mahjong puzzle, consists of establishing a correspondence between shuffled tiles.

In this exercise, the objective is to match words with their definitions. During the process of building each puzzle, a filter chain is used to select the word-definition pairs. This is done to automatically exclude long definitions, similar words (in terms of concept) and characters that might affect the comprehension of the exercise (like punctuation elements, the cardinal character, and others).

After this step, a set of classifiers is used to establish the degree of difficulty and whether certain definitions are linked to specific domains. This is due to certain words being homonyms which makes certain definitions to be considered advanced/not trivial to everyone. This process is done once for every word from the puzzle and its results are stored in the REAP.PT database.

Finally, like the original Mahjong puzzle, a scoring mechanism was also incorporated. Since the final scored is calculated using both the elapsed time and the number of mistakes made during the exercise resolution, it serves as a useful tool for both teacher and student to assess the student's performance.

An example of this exercise can be found in Figure 2.3.

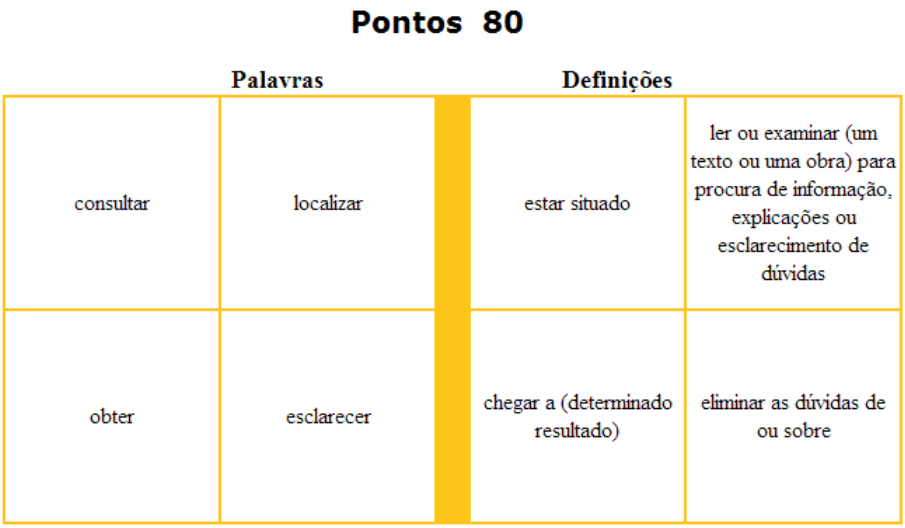

Figure 2.3: REAP.PT lexical mahjong exercise.

2.2.2 Collective Names and Nominal Determinants

The collective names and nominal determinants exercises were developed by Marques (Marques, 2011) and their goal is to teach students two different aspects regarding nouns. The constraints between determinative nouns and the nouns they determine, for example *um cesto de fruta* (a basket of fruit), and the relations between common nouns and their respective collective nouns, for example *um rebanho de cabras* (a herd of goats). The latter occurs when the collective noun is also a determinative noun.

Both exercises use sentences extracted from the REAP.PT corpus which are later processed by STRING (Mamede et al., 2012), the natural language processing chain of L^2F. During this stage, a quantifying dependency is established between the phrase (nominal or prepositional) that contains the determinative noun and the consecutive prepositional phrase that contains the determined noun.

Following the identification of the determinative noun, the next step consists of generating distractors. These share the gender and number with the extracted noun.

"Sarcastic" relationships between nouns that could inadvertently mislead the user into selecting a wrong answer are purposely avoided. For example, it is impossible to generate distractors that contain a determinative noun with an associated "Animal" feature and a determined noun with an associated "Human" feature (e.g. *alcateia de políticos*, a pack of politicians). Finally, generic determinative nouns (e.g. *grupo*, group) are also avoided.

If the student chooses the wrong answer when solving the exercise, a feedback system consisting of three different components appears. It consists of: a definition of the wrong answer, an example featuring the chosen answer and finally, a picture of the given answer. These images were previously stored in the REAP.PT database.

The Collective Names exercise is based on a specific subset of questions generated in the Nominal Determinants exercise. Correct answers consisting solely of collective names, are filtered and used for this purpose. Separate exercises were designed in order to make the process of learning what a collective name refers to, easier.

Examples of these exercises can be found in Figures 2.4 and 2.5.

2.2.3 Choice of Mood in Subordinate Clauses

The choice of mood in subordinate clauses exercise was also developed by Marques (Marques, 2011). In this exercise, the objective is to select the correct verb, from several alternatives, that fits the syntactic restrictions that connect the main clause of a given sentence and the subordinate clause where the verb form to insert is missing. These are typically related to the tense and mood of the main clause.

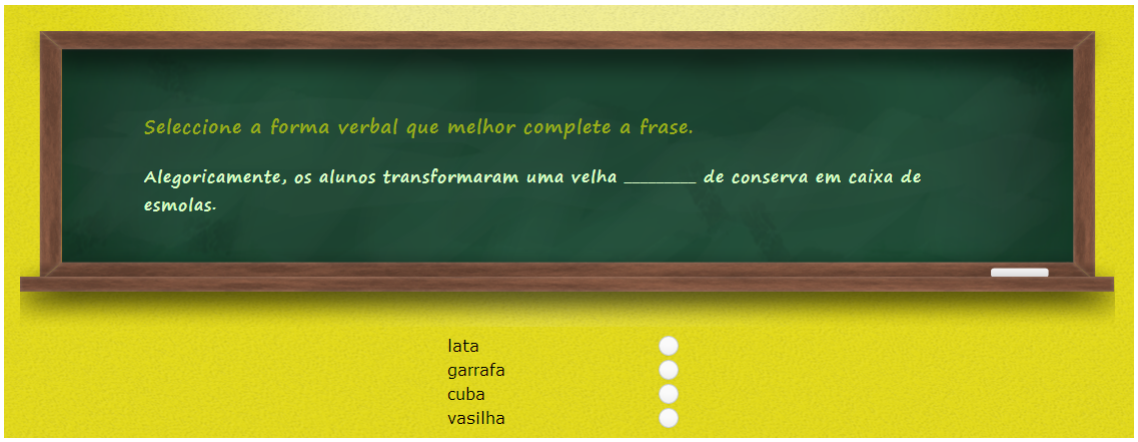

Figure 2.4: REAP.PT nominal determinants exercise.

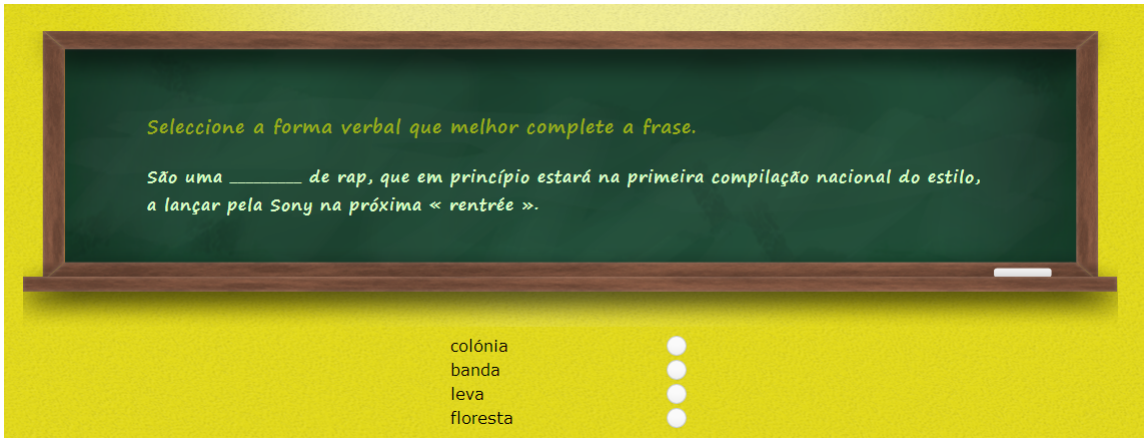

Figure 2.5: REAP.PT collective names exercise.

To generate this exercise, sentences are first extracted from CETEMPúblico (Santos & Rocha, 2001), a corpus consisting of texts extracted from 2.600 editions of the Público Newspaper, and then processed by STRING. It is during this step that the links between the clauses are established and the verb of the subordinate clause is identified.

The L^{2}F VerbForms[3] generator is then used to automatically generate the wrong answers (also know as distractors), based on the previously identified verb. To avoid ambiguity and therefore, more than one correct answer, a set of rules is applied during this step.

An example of this exercise can be found in Figure 2.6.

[3] `https://www.l2f.inesc-id.pt/wiki/index.php/VerbForms` (last visited in September 2013)

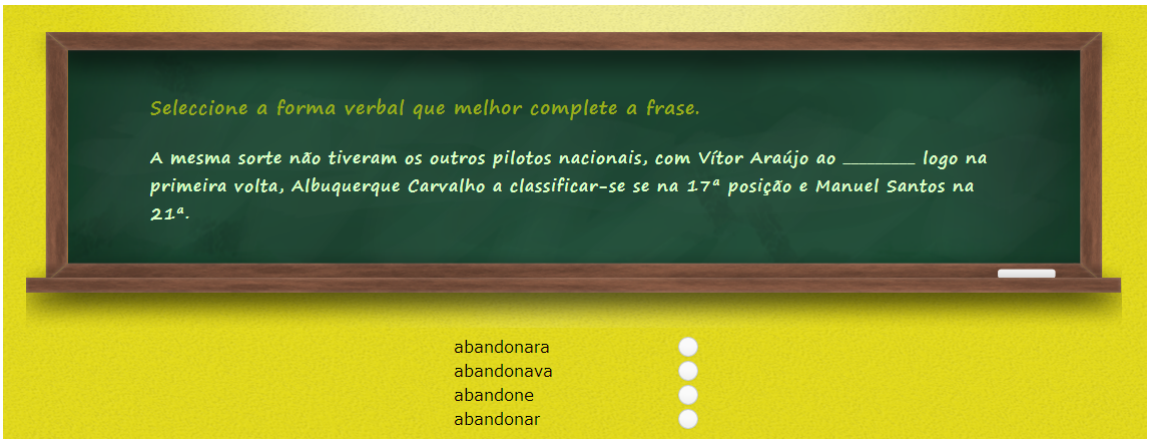

Figure 2.6: REAP.PT choice of mood in subordinate clauses exercise.

2.2.4 Pronominalization

The pronominalization exercise was developed by Freitas (Freitas, 2012) and its goal is to teach students how to replace a constituent with a pronoun in a given sentence. This is typically done automatically by speakers in their everyday conversations when referring multiple times to a certain subject. The constituent in this case can be a prepositional phrase, a noun phrase or a substantive subordinate clause. Each of these has its own restrictions when it comes to replacing it with a pronoun and placing said pronoun next to the verb of the sentence.

When generating this exercise, sentences without pronouns are selected from the REAP.PT corpus. After the selection stage, a series of filters is applied to exclude longer sentences and complex sentences that already contain the desired answer for the exercise, elsewhere in the sentence. For example, in the sentence *O Pedro encontrou a Ana e cumprimentou-a alegremente.* (Pedro met Ana and happily greeted her.), the pronoun *a* not only refers to *Ana* but, it is also the intended pronoun for the constituent's replacement.

Following this, additional filters are applied depending on the form of the constituent. After this, the sentence is processed by STRING, to identify the complement dependencies between phrases. It is also in this stage that the gender and the number of the pronoun are identified. These aspects are used when generating the distractors that will be presented to the user. Between these stages, a set of rules specify where the pronoun is placed. This typically occurs before or after the verb, which can be the either the auxiliary or the main verb.

Finally, the distractors are generated according to a series of parameters, which typically involve the incorrect positioning of the pronoun, an incorrect identification of the pronoun's case or both.

An example of this exercise can be found in Figure 2.7.

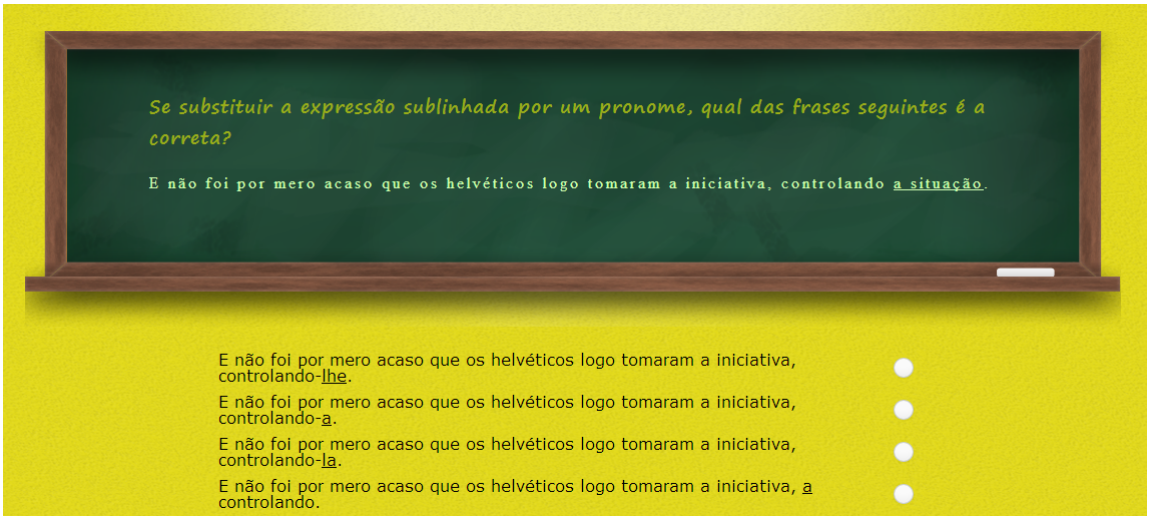

Figure 2.7: REAP.PT pronominalization exercise.

3 State of the Art

In this section, several Portuguese CALL systems which share some features with REAP.PT are presented.

3.1 Portuguese CALL Systems

3.1.1 Ciberescola da Língua Portuguesa

Ciberescola da Língua Portuguesa[1] is an online platform for teaching Portuguese. It is different from some of the other available CALL systems because it aims to teach both native (grades 5th through 12th) and foreign students (levels A1 through C2). Because it is owned by a non profitable company, access to its contents is completely free.

It currently offers 1000 exercises, covering various topics (reading, oral comprehension, grammar, writing and vocabulary), sorted according to the user's school grade and the desired degree of difficulty (easy, normal and hard). Different types of exercises are available, including, but not limited to true/-false questions, fill-in-the blanks and multiple choice questions. Multimedia resources, such as videos, pictures and audio clips, are often presented beforehand, to the users, to help contextualize the exercises.

Being interactive exercises, students receive instant feedback upon completing these. Following the identification of the correct and incorrect answers, the student is presented with the achieved score and the answer to the exercise. The web-page's interface also allows the user to keep track of the exercises already completed and their respective score.

However, if the user incorrectly solves an exercise, there is no additional feedback to help him/her improve his/her language skills or to correct the cause of the mistakes. Also, contrary to what happens with REAP.PT, the exercises in Ciberescola are not automatically generated. However, they are entirely produced and managed by teachers and researchers in key areas, namely linguistics, language teaching and literature.

[1] `http://www.ciberescola.com` (last visited in September 2013)

Figure 3.1: Ciberescola Web-page - Odd man out exercise.

3.1.2 Centro Virtual Camões

Centro Virtual Camões[2] is part of the Instituto Camões, a public institute whose mission is "to plan and implement the policy for the promotion, dissemination and teaching of the Portuguese language and culture abroad and to promote Portuguese as a language of international communication"[3].

The exercises currently provided by this web-page focus on the reading and listening aspects of the Portuguese language. All of them are distributed among three difficulty levels. Other resources also focus on the speaking part of the language but these do not include any exercises.

The reading exercises, provide the users with a series of texts and multiple-choice questions based on these. In some cases, the user is only allowed to progress if the correct alternative is selected. If the opposite occurs, an explanation and description of the error is presented to the user. The listening exercises make use of recorded audio clips and provide true/false, sorting and sequencing questions based on these. Some of these clips consist of book excerpts from selected Portuguese authors.

Besides these, there are also traditional playful exercises available such as the Hangman Game and the Glory Game.

Like the exercises from Ciberescola, the ones provided by Centro Virtual Camões are also not auto-

[2] `http://cvc.instituto-camoes.pt/aprender-portugues.html` (last visited in September 2013)
[3] `http://www.instituto-camoes.pt/informacao-institucional/missao-do-camoes` (last visited in September 2013)

Figure 3.2: Centro Virtual Camões Web-page - *Roda da Fortuna* (Listening exercise).

matically generated, neither are they tailored to the user's topics of interest.

3.1.3 Observatório da Língua Portuguesa

Observatório da Língua Portuguesa[4] is a non-profit organization dedicated to promoting the teaching of the Portuguese language.

Even tough some exercises are available on their web-page, the main focus of this platform is to offer reports and other data, on topics ranging from the economical value of the Portuguese language to its projection and utilization throughout the world, regarding the number of speakers. Several newspaper articles and statistical reports regarding these subjects are available.

The available lexical exercises, with the majority of them being multiple choice or fill-in-the-blank questions, are implemented using Google Drive Form sheets and thus, they are not automatically generated. Also, no correction is provided after the answers are submitted. Finally, unlike REAP.PT, the exercises are not specifically tailored to the users' areas of interest.

[4] `http://observatorio-lp.sapo.pt` (last visited in November 2012)

Figure 3.3: Observatório da Língua Portuguesa - Fill-in-the-blanks exercise.

3.2 Foreign ICALL Systems

In this section, several foreign ICALL systems are presented. Similar to what was done in the previous section, the chosen systems were the ones sharing features with REAP.PT.

3.2.1 TAGARELA

TAGARELA (Teaching Aid for Grammatical Awareness, Recognition and Enhancement of Linguistic Abilities)[5] (Amaral & Meurers, 2011) is an ICALL system designed for the teaching of the Portuguese language. Development began under the ICALL Research Group[6] at Ohio State University, and further improvements were made at Tübingen University's Department of Linguistics[7].

According to the information available at the site[8], the system employs Natural Language Processing to analyse the user's input and identify errors, namely, of the semantic, syntactic, morphological and spelling type. The exercises provided are similar to the ones found in common textbooks and they include listening and reading comprehension, picture description, vocabulary practice, phrasing and re-writing.

It shares many of REAP.PT's features, namely, the ability to keep track of the user's performance throughout the solved exercises and to provide personalized feedback based on the specificity of the

[5] `http://purl.org/icall/tagarela` (last visited in September 2013)
[6] `http://www.ling.ohio-state.edu/icall` (last visited in September 2013)
[7] `http://www.sfs.uni-tuebingen.de` (last visited in September 2013)
[8] `http://sifnos.sfs.uni-tuebingen.de/tagarela/index.py/main?what=about` (last visited in September 2013)

errors and the exercise at hand.

The exercises were unavailable at the time of this document's elaboration.

3.2.2 WERTI

WERTI (Working with English real-texts)[9] (Meurers et al., 2010) is an ICALL system designed to help English as a second language students overcome their shortfalls. Like the TAGARELA system, development of WERTI began at Ohio State University, under the ICALL Research Group and was continued later at Tübingen University.

WERTI shares many of REAP.PT's features. First, it also uses real texts, extracted from the web, to generate its exercises. Second, this step is done by employing Natural Language Processing to identify specific resources within these texts, such as parts of speech or lexical materials. However, unlike REAP.PT, WERTI does not use previously filtered web-pages. Instead, the users are prompted to install a toolbar plugin in their browser (only Firefox is supported), leaving the choice of the web-page to be processed, up to themselves. Because of this, only the NLP part of this processing chain is done at the server's side.

Multiple types of exercises are available for a specific set of topics with the latter consisting of articles, determiners, gerunds, noun countability, phrasal verbs, prepositions and *wh*-questions. Currently, the system provides clicking, multiple-choice and fill-in-the-blanks exercises. Regardless of the topic of interest and exercise type chosen by the user, almost any combination is available. The clicking exercise requires the user to click all elements of a specific topic in the current page. The multiple choice exercise replaces the elements with a drop-down menu containing the correct answer and automatically generated distractors. Finally, the fill-in-the-blanks exercise (referred to in the toolbar as "practice") also replaces the elements but this time, a white box is provided for the user to input the correct answer. In all three exercises, feedback is provided by colouring the users' answers. Correct answers are coloured green and incorrect answers are coloured red.

Apart from these exercises, the plugin also provides an additional colourize operation that highlights each word belonging the select topic. Support for the German and Spanish languages is also available albeit at a Beta stage.

[9] `http://sifnos.sfs.uni-tuebingen.de/WERTi/` (last visited in September 2013)

3.2.3 Alpheios

The Alpheios Project[10] is an open-source project developed by The Alpheios Project non-profit organization.

This project focuses primarily on classical literature and dead languages, such as Latin and ancient Greek however, in the long run, resources for other languages, including modern ones will also be made available.

Like WERTI, this project uses a Firefox toolbar plugin that allows the user to perform several operations on selected texts. These include classic works such as the Iliad by Homer and the One Thousand and One Nights. When the texts are presented to the user, one of two modes (reading mode or quiz mode) must be chosen by the user.

In the first one, not only is an input field made available for looking up the morphological analysis and translation of words but, by double-clicking any word from the text, a window, including the translation of the given word and additional operations, pops-up. In this pop-up, the user can perform several operations, such as looking up the definition of a the word in the dictionary, viewing a tree that represents the syntactic analysis of the sentence containing the word or consult the word's inflection tables and mark it as "learned". The latter adds the word to a list the user can access at any time and that also features the words he or she double-clicked. Greek texts also feature a translation panel that allows the user to view both the original and translated text side-by-side.

In the quiz mode, the user is required to double-click any chosen word and is then presented with a different question depending on the language of the text. In Latin texts for example, the user is asked to find the matching word in the translation panel. In Greek texts, the user is presented with a pop-up containing a multiple-choice question regarding the the part-of-speech. The choices are fixed and encompass all possible parts-of-speech so, specific distractors based on the word are not automatically generated. Exercises with Arabic texts are not yet available.

Regardless of the selected mode, additional resources for the Greek and Latin languages namely, inflection tables and grammars are always available from the toolbar. The user can also manually enter texts in and view the corresponding syntactic trees.

3.3 Current Syntactic Exercises on Word Formation

In this section, examples of exercises on the topic of word formation, found in Portuguese textbooks, are presented. These will focus on small variations of the exercises proposed in this project.

[10] `http://alpheios.net/` (last visited in September 2013)

3.3.1 Syllable Division

Exercises focusing on syllable division typically require the student not only to perform said division, but also to analyse the words that are presented to him, according to certain parameters. This process usually revolves around categorizing words according to their number of syllables or according to the position of their tonic syllable. Some examples of these exercises are presented below.

1. Example adapted from *Gramática da Língua Portuguesa (2.° Ciclo do Ensino Básico 5.°/6.° Anos)* (Lopes, 2011); this exercise tests several rules of syllable division such as the maintenance of the digraph *lh* or the splitting of the digraph *ss*:

   Faz a divisão das sílabas das palavras apresentadas, de acordo com o modelo. (Divide the words into syllables according to the example).

   espetáculo (spectacle) → *es-pe-tá-cu-lo*

   (a) *floresta* (forest)
   (b) *escuteiro* (scout)
   (c) *espelho* (mirror)
   (d) *assombro* (astonishment)
   (e) *glamoroso* (glamorous)

2. Another example from the same textbook; this exercise reinforces the notion of syllable by its reuse in forming different words:

   A partir da sílaba destacada em cada uma das palavras e, de acordo com o modelo, forma uma nova palavra. (Form a new word, according to the example, using the syllable in bold).

   *infor**má**tica* → ***má**goa*

   (a) *fol**ga**zão* (merry)
   (b) ***pro**grama* (program)
   (c) *senti**men**to* (feeling)
   (d) ***al**vorada* (dawn)
   (e) *sau**da**de* (nostalgia)

(f) *a**le**gria* (joy)

3. A final example adapted from *Domínios-Gramática da Língua Portuguesa, Exercícios de Aplicação* (Lopes, 2012); this exercise aims at classifying the words according to the position of their tonic syllable:

   Preenche o quadro utilizando as palavras fornecidas. (Fill the table using the available words).

Palavras Agudas	*Palavras Graves*	*Palavras Esdrúxulas*

   Lista de Palavras: corpo (body), *escura* (dark), *maldição* (curse), *morrer* (die), *zoológico* (zoological), *mares* (seas), *veterinários* (veterinarians), *tentou* (tried), *substância* (substance).

3.3.2 Derivation

Regarding word formation, there are several types of exercises available. In order to present alternatives to the proposed exercises, the search will be narrowed to the ones focusing on derivation by prefix addition.

Various examples can be found in Portuguese textbooks:

1. Example from *Gramática da Língua Portuguesa (2.° Ciclo do Ensino Básico 5.°/6.° Anos*) (Lopes, 2011); this exercise aims at identifying which words are formed through prefixal derivation:

   Assinala na lista duas palavras derivadas por prefixação (Identify two words formed by prefixal derivation):

   (a) *laranja* (orange)
   (b) *régua* (ruler)
   (c) *pente* (comb)
   (d) *rever* (review)
   (e) *luz* (light)
   (f) *transpor* (transpose)
   (g) *vassoura* (broom)
   (h) *lápis* (pencil)
   (i) *folha* (sheet)
   (j) *livro* (book)

2. Another example from the same textbook; in this exercise, the goal is to apply prefixes to negate the given verbs:

   Indica o contrário de cada um dos verbos apresentados, empregando prefixos (Apply prefixes to find the opposite of the following verbs):

   (a) *aparecer* (to appear)
   (b) *aconselhar* (to advise)
   (c) *viabilizar* (to make possible)
   (d) *obedecer* (to obey)
   (e) *disponibilizar* (to make available)
   (f) *organizar* (to organize)

3. A final example from *Domínios-Gramática da Língua Portuguesa, Exercícios de Aplicação* (Lopes, 2012); this exercise aims at correctly matching prefixes and words:

   Das formas de prefixo seguidamente apresentadas, seleciona aquela que possas adicionar a uma das palavras abaixo grafadas. (For each of the given words, choose an applicable prefix).

   i- im- des- re- ex- pre- pré- pro- super- sub- tres- trans- ultra- anti- hiper- di-

   (a) *ativo* (active)
   (b) *emprego* (job)
   (c) *potência* (power)
   (d) *visão* (vision)
   (e) *passar* (pass)
   (f) *começar* (begin)
   (g) *capaz* (capable)
   (h) *aquático* (aquatic)
   (i) *violeta* (violet)
   (j) *mercado* (market)
   (k) *fuga* (escape)
   (l) *ministro* (minister)
   (m) *história* (history)
   (n) *coagulante* (coagulant)
   (o) *legível* (legible)
   (p) *sílaba* (syllable)

4 Architecture

In this section, the architecture of the syllabic divisor and both exercises is described.

4.1 YAH

Since both exercises focus on word formation, there was the need to automate the process of dividing words into syllables. Originally, the plan was to use the syllable-divided vocabulary of P-AWL in both exercises, however, given the size and the scope of the alternative available resources (Vicente, 2013) for the Derivation exercise, the inclusion of the words belonging to the P-AWL (Baptista et al.,), in this specific exercise, was dropped. The Crossword Puzzle exercise, on the other hand, only uses words which are present in the P-AWL.

Furthermore, there are different interpretations of how words should typically be divided. Among these, are the processes of dividing words phonetically[1] and orthographically. The latter, which corresponds to the act of hyphenating words, is done according to the rules in the *Acordo Ortográfico da Língua Portuguesa de 1990*[2] (Portuguese Language Orthographic Agreement of 1990). In this exercise, this specific set of rules was followed, albeit with some changes. These concern the division of the digraphs "ss" and "rr", in view of the specific goal of the exercises.

During the hyphenation process of words containing one of these digraphs, such as *possibilidade* (possibility), the orthographic result is the splitting of said digraph (*pos-si-bi-li-da-de*). However, if we ask any native speaker how he or she would utter this word, syllable by syllable, since this digraph constitutes a single phone [s], both "s's" would be placed right next to each other in the second syllable, with the transcription of this division being *po-ssi-bi-li-da-de*. Therefore, in view of the phonological goal of the exercise, which is to identify syllables we consider that this is the best approximation, even though not orthographically correct.

Concerning the digraph "rr", the same problem occurs in words such as *carregado* (loaded) or *barro* (clay). The result of orthographically hyphenating these words is *car-re-ga-do* and *bar-ro* respectively,

[1] `http://www.academia.org.br/abl/cgi/cgilua.exe/sys/start.htm?infoid=11762&sid=784` (last visited in August 2013)

[2] `http://www.portaldalinguaportuguesa.org/acordo.php?action=acordo&version=1990` (last visited in August 2013)

however dividing these words into syllables, according to the previously specified parameters, results in *ca-rre-ga-do* and *ba-rro*.

With this in mind, in order to prevent mistakes and to simplify the division process, an hyphenator entitled YAH[3], which stands for "Yet Another Hyphenator", was used as the basis for the task of dividing the P-AWL. This C++ implemented tool, which had already been developed at L^2F, receives a word as its input parameter and produces an hyphenated string as its output. This is done by applying a "mask" to the word, which classifies each letter according to its type (consonant or vowel).

Tables 4.1 and 4.2, illustrate the differences in the masking process of both versions of the YAH hyphenator, the original version and the modified and expanded version, used in this project.

Mask Character	**Characters**
c	bcçdfghjklmnpqrstvxz
p	ptqbdgfvc
l	lr
v	aáàãâeéèêiíîoóòõôuúùû

Table 4.1: Original YAH mask characters.

Mask Character	**Characters**
c	bcçdfghjklmnpqrstvxz
p	ptqbdgfvc
l	lr
v	aáàãâeéèêiíîoóòõôuúùû
i	eéèêiíî
u	oóòõôuúùû

Table 4.2: Modifed YAH mask characters.

Assignment of the consonant characters in both versions and vowel characters in the modified YAH, is done with a simple If-Then-Else clause. For example, given the character **b**, the system tests whether it is a consonant or a vowel. Since this character is a consonant, another test is made to check whether it is a "p" type consonant or a "l" type consonant. Given the fact that it belongs in the first category, the "p" mask character is assigned to it.

After this step, the input string is matched against a set of "word beginning rules", which as the name suggests, concern the first letters of the word. Whenever there is a positive match, the first matching rule is applied. This process repeats itself throughout the hyphenation of the whole word, with the

[3]The hyphenator has been built by Ricardo Ribeiro, in 2012.

remainder of the applicable (generic) rules. It is important to note that even though the existence of the mask generalizes the division process, it is also possible to create and apply rules depending on a specific sub-string of characters.

Some of these rules are presented in Figure 4.1.

```
word_beginning_rules.push_back(Rule("reo", 2));
word_beginning_rules.push_back(Rule("sta", 3));
word_beginning_rules.push_back(Rule("subl", 3));
```

Figure 4.1: YAH word beginning rule examples.

For example, the third line of code concerns words beginning with "subl", such as *sublinhar* (underline), *sublime* (sublime) or *subliminar* (subliminal). It states that an hyphen should be placed at the third position of the word (hence the number 3), producing a temporary output string consisting of "sub-l".

After one of these beginning rules is applied or if none of them match the beginning of the word, a set of more generic rules is then used. This is where the main modifications were made, in order to achieve the desired results. The original YAH contained only eight word beginning rules and eight more rules which did not guarantee the correct division of most of the words present in the P-AWL. The final version of the modified YAH contains sixteen word beginning rules, fifty six "generic" rules and thirteen exceptions that deal with word endings and other peculiar aspects. The complete listing of all the rules can be found in Annex A. The latter were implemented to deal with certain syllables which are processed by the existing rules (because a valid mask is applied) but constitute exceptions which should be dealt with accordingly.

```
pos = result.find("tui=");//exception
if (pos != string::npos)
{
      result.replace(pos, 4, "tu=i=");
}
```

Figure 4.2: YAH exception example.

For example, the code in Figure 4.2, deals with the way the syllable "tui" is read, whenever it is not the final syllable of the word. Instead of the diphthong "ui" present in "*constitui*" (constitutes), this string could represent a hiatus as it happens in the word "constituição" (constitution).

In order to assess the overall quality of these rules, an alternative corpus consisting of 1722 words was created. This was done by extracting the contents of eight random on-line articles from the Público newspaper[4], as well as 126 words taken from the examples of the hyphenation rules, provided by the

[4] `http://www.publico.pt/` (last visited in September 2013)

Acordo Ortográfico da Língua Portuguesa de 1990. The modified YAH hyphenator was capable of correctly dividing 97% of this corpus, which corresponds to 1671 correctly divided words out of the total 1722. In comparison, the original YAH was only capable of dividing 74,2% of this corpus contents and 63,1% of the P-AWL. Regarding the modified YAH, the few cases which prevent a 100% rate from being achieved, mainly concern design choices which were taken with the specific contents of the P-AWL in mind. For example, the word *"reino"* (kingdom), is incorrectly divided as *"re-ino"* when it should be divided as *"rei-no"*. This happens because all the words in the P-AWL beginning with the string "rein", are formed by adding the prefix "re-" to an existing form, such as *"reincidir"* (reoccur), *"reinterpretar"* (reinterpret) or *"reinvestir"* (reinvest). Being formed by prefix addition means that in this case, when hyphenating the words, the prefix corresponds to the first syllable. In *"reino"*, the substring *"ei"* corresponds to a diphthong which should not be divided.

Besides not being able to cover all the aspects concerning the hyphenation of Portuguese words, such as the ascending diphthongs (where a semi-vowel precedes a vowel), descending diphthongs (the opposite), and the digraphs "qu","gu", "pn" or "pt", among others, which are all indivisible, the original YAH was not capable of processing words containing letters with diacritical marks, such as "ç" or "á". The latter problem was solved by modifying the way each word was parsed and how the mask was applied. The remaining aspects were corrected by implementing additional rules.

In Table 4.3, the differences in the masking and hyphenation operations, between both systems, can be observed.

Word	**Mask**	**Hyphenization**	**Tool**
concentração	`pvcpvcplvv?v?v`	`con=cen=tração`	Original
	`pucpicplvpvu`	`con=cen=tra=ção`	Modified
ambientalistas	`vcpvvcpvlvcpvc`	`am=bien=ta=lis=tas`	Original
	`vcpiicpvlicpvc`	`am=bi=en=ta=lis=tas`	Modified
atribuí	`vplvpvv?`	`a=tri=buí`	Original
	`vplipui`	`a=tri=bu=í`	Modified
multidimensional	`cvlpvpvcvccvvcvl`	`mul=ti=di=men=sio=nal`	Original
	`culpipicicciucvl`	`mul=ti=di=men=si=o=nal`	Modified

Table 4.3: Comparison of the original and the modified YAH hyphenators.

The question mark characters symbolizes the diacritical mark, which the application does not know how to interpret, that is, whether it is a consonant or a vowel. This occurs because letters with diacritical marks when coded, are actually composed of two consecutive elements: the letter itself and the diacritical mark. By looking ahead to see if the next element, in the string corresponding to the word, is in fact a diacritical mark, one can process these characters accordingly. This prevents errors such as the one concerning the word *concentração* (concentration), in which all the letters with diacritical marks are

treated as a combination of a vowel and a question mark. Since there are no rules capable of processing masks with questions marks in them, these words were not correctly divided.

Furthermore, the mask character symbolizing a vowel (v) was expanded into several different characters in order to deal with the issues regarding the diphthongs. This allowed for more complex rules to be implemented, such as the one in in Figure 4.3.

```
rules.push_back(Rule((std::string(1,consonant_mask)+=
                                    type_i_vowel_mask)+=
                                    type_u_vowel_mask, 0, true));
```

Figure 4.3: YAH generic rule example.

This specific rule states that the words with a matching "ciu" mask, meaning any consonant from "hjkmnsxz", followed by an "i" type vowel and an "u" type vowel, such as *multidimen**sio**nal* (multidimensional) or *indu**ziu*** (induced), are hyphenated at the beginning of the matching set of characters (hence the zero). Moreover, this rule belongs to a subset of rules which are processed "twice", meaning that another hyphenation mark is inserted in another position. This is done elsewhere in the code by identifying which rule has been matched and applying another hyphenation mark to a different position, which in this case is done between the "i" vowel and the "u" vowel. In the end of this process, the "ciu" mask is modified and the result is another mask which is then applied to the word in order to insert the hyphenation marks in the correct places. In this specific example, the result would be a new mask consisting of "-ci-u", which when applied to the words mentioned before would hyphenate them as *mul-ti-di-men**-si-o**-nal* and *in-du**-zi-u***

It is important to note that the purpose of these modifications was not to create an hyphenator with the best success rate possible, but to create a tool capable of correctly dividing the entirety of the P-AWL's contents, that is, its 2553 distinct word forms. With that being said, once the P-AWL was correctly divided, no extra modifications were made in order to cover a bigger set of words and achieve an even greater range of "divisible" words. However, the positive results, obtained from hyphenating the contents of the alternative corpus, prove that the YAH hyphenator can successfully cover a bigger variety of words than those which compose the P-AWL.

4.2 Crossword Puzzle Exercise

Portuguese Crossword Puzzles are normally filled using letters. However, separating words into syllables is also an useful exercise when the objective is to teach both vocabulary and diction, since analysing the syllables that compose a word and knowing how to pronounce them separately, makes the task of reading words aloud an easier one. In this exercise, the goal is to solve a crossword puzzle grid by

dividing words into syllables. The welcome web-page can be found in Figure 4.4 and an example of the exercise can be found in Figure 4.5.

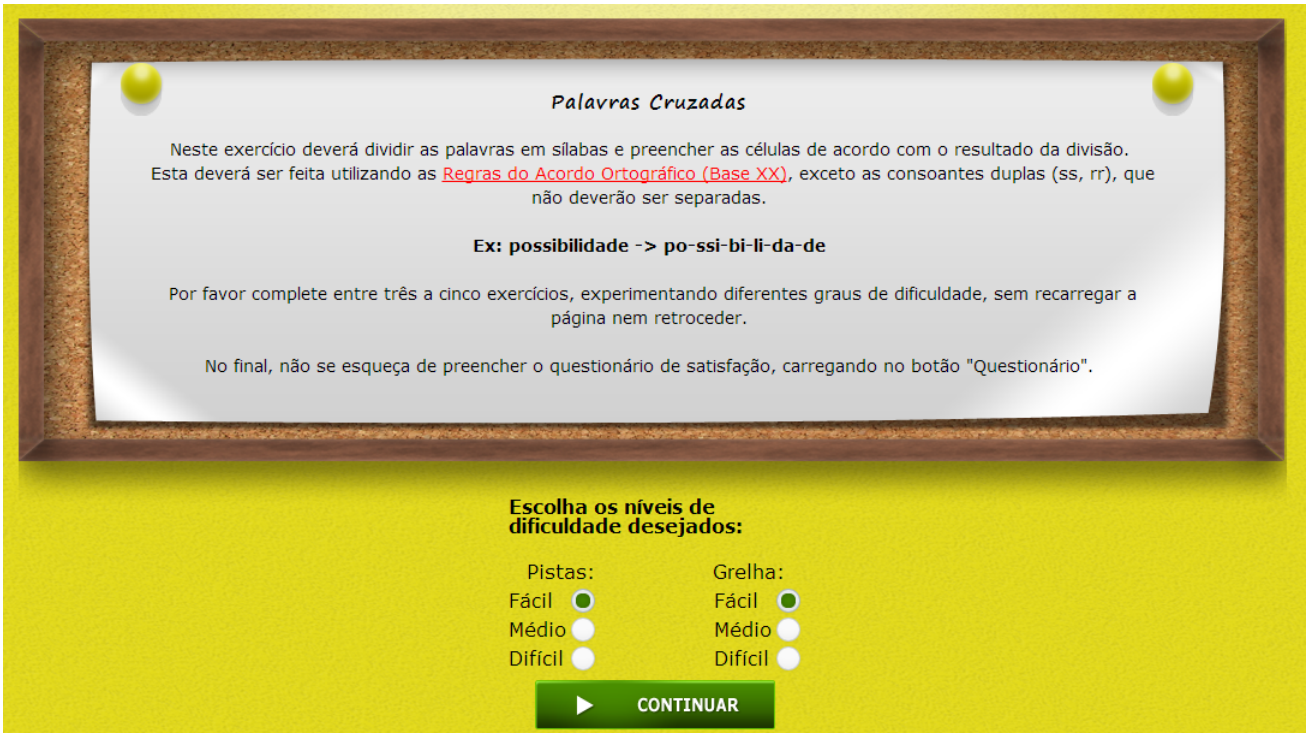

Figure 4.4: Crossword exercise welcome web-page.

Splitting words into letters can be considered a trivial task, since letters are the smallest distinguishable unit in most written languages. However, splitting words into syllables requires certain rules to be followed, which, of course, differ from language to language. Regarding Portuguese, in general, and this exercise in particular, the modified YAH hyphenator was used to divide the P-AWL in its entirety. The result of this operation is then used as the corpus from where the divided words are extracted and used to build the puzzle. This corpus consists of 570 lemmas and 1983 distinct inflected forms, for a total of 2553 words (most inflected forms were not considered).

Figure 4.6 shows the steps that are taken in order to generate each puzzle.

4.2.1 Word Selection

In the first step, which corresponds to the index page of the exercise, the user is asked to choose between three difficulty levels for two different parameters: the grid and the definitions (hints).

After this, one word is randomly chosen from the previously divided P-AWL, and used as the target vertical word. This is the only word that has to make sense vertically. In order to prevent the generation

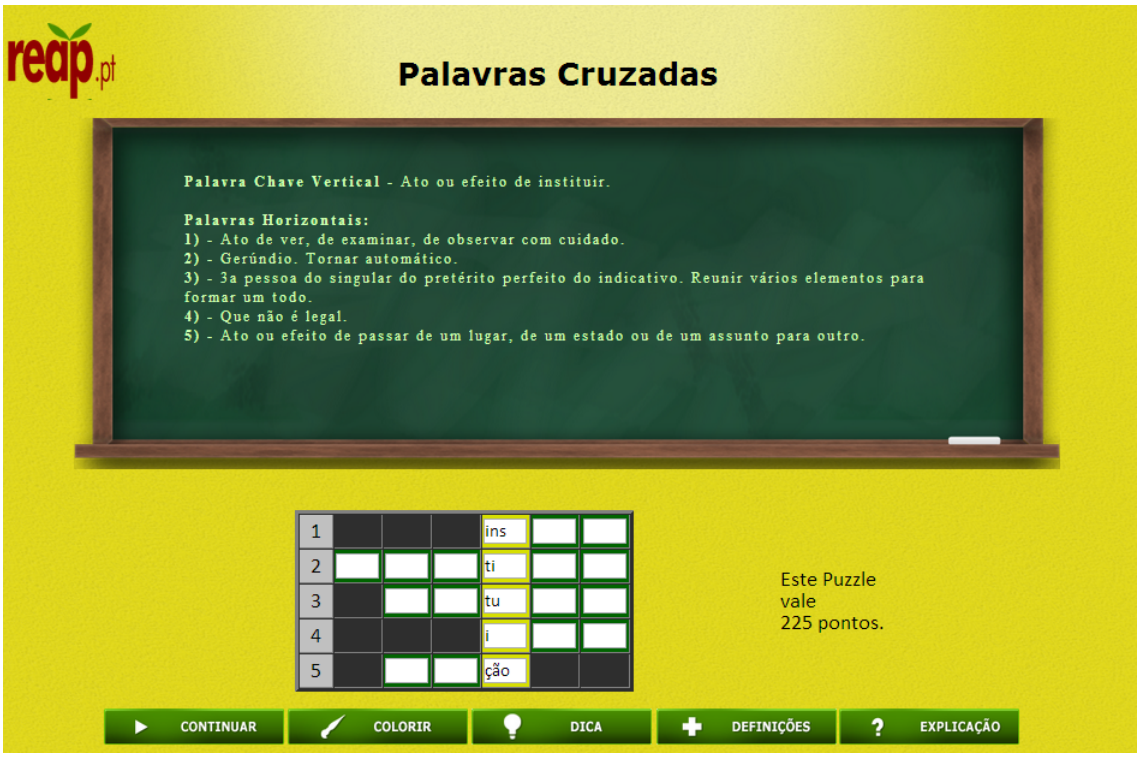

Figure 4.5: Crossword puzzle exercise interface.

of small grids, in this phase, words with less than four syllables are rejected, which means that from the 2553 total words, only 1665 are eligible for this process.

Since the selected word is already divided into syllables, these are separately used in order to find the matching intersections for the horizontal words. Once again, these are extracted from the P-AWL.

In order to obtain a more aesthetically pleasing puzzle, words beginning with the target syllable have priority over words containing the syllable in another position. For example, if the vertical word is *implicitamente* (implicitly), which is divided as *im-pli-ci-ta-men-te*, a word such as *mentalidade* (mentality) would be chosen as the fifth horizontal intersection over another word such as *acompanhamento*

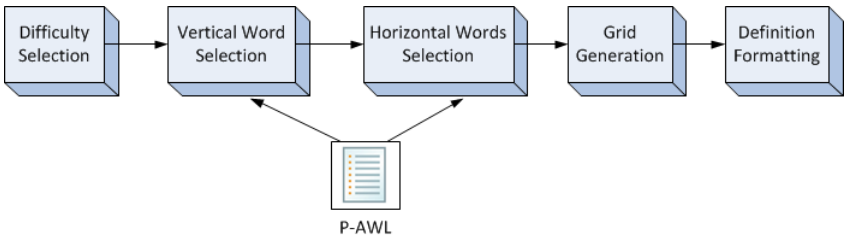

Figure 4.6: Crossword puzzle generation steps.

(accompaniment).

To prevent similar words being chosen, the Levenshtein distance (Levenshtein, 1966) between the vertical word and each of the intersecting words, is calculated. The simplest PHP implementation[5] of this function was used, which means that the deletion, insertion and replacement operations all have the same cost, which is one. With this in mind, if the result of the comparison between the vertical word and the intersecting word is less than five, the intersecting word is rejected. This threshold value prevents the selection of similar words such as *reestruturada* (restructured) and *reestruturação* (restructuring).

4.2.2 Grid Generation

After the completion of the last step, a fully filled grid is generated. This is used as a basis for the grid which the user will be asked to fill. Using the previously selected difficulty level, several transformations are made to this grid.

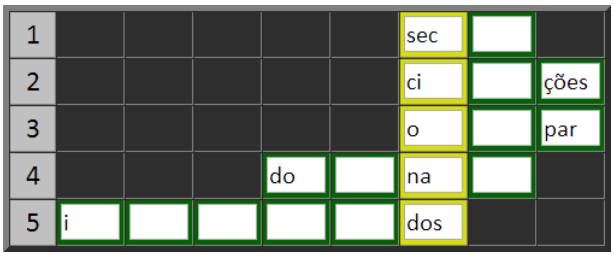

Figure 4.7: Example of an easy Crossword puzzle grid.

An example of an easy puzzle grid can be found in Figure 4.7. If the user selects this level (easy), the first and last cell of each word are left correctly filled as well as the cells corresponding to the vertical word. However, if these cells are adjacent, that is, if they are immediately next to each other, the cell corresponding to the beginning or the end of the intersection, is deleted. This is done to increase the challenge of solving each puzzle. For example, in the example of Figure 4.7, the vertical word is *seccionados* (sectioned) and the fourth intersection is *dominação* (domination). However, only the syllable *do* appears in its cell, in the grid, because even tough *ção* corresponds to the word's last syllable, it is adjacent to the vertical word's syllable, and thus, it is removed from the grid.

On the other hand, if the user selects the medium difficulty level, only the vertical word's cells are left filled in the grid. Finally, if the hard level is chosen, the grid's contents are completely erased, leaving nothing but blank, empty cells.

[5] `http://php.net/manual/en/function.levenshtein.php` (last visited in August 2013)

4.2.3 Definition Formatting

The definitions of the words in P-AWL were previously extracted using both the *Infopédia*[6] and Priberam[7] on-line dictionaries, with the latter being used for adverbs, since the definitions in the Infopédia dictionary were considered unhelpful. However, the results of these extractions need to be formatted in order to be presented as hints to the user.

• moedas: nome feminino plural de *moeda*

moeda

nome feminino

1. peça geralmente metálica, cunhada com autorização legal, que serve para realizar transações financeiras
2. unidade monetária em vigor em determinada região
3. tudo o que representa um valor pecuniário
4. ECONOMIA instrumento que exerce as funções de meio de pagamento, medida de valor e meio de reserva
5. ECONOMIA conjunto dos meios de pagamento de uma economia (moeda metálica, papel-moeda, moeda bancária)
6. antiga moeda portuguesa, do valor de 4800 réis
7. *figurado* aquilo que serve como meio de troca

moeda corrente/sonante

dinheiro em metal, em circulação num país, por oposição a papel-moeda

moeda fiduciária

moeda em papel que não está garantida por igual valor em metal, circulando com base na confiança

moeda forte

moeda com elevada aceitação, dada a sua estabilidade internacional

moeda fraca

moeda com reduzida aceitação, devido à sua tendência para a depreciação

moeda única

unidade monetária (euro) adotada pela maioria dos estados-membros da União Europeia, em substituição das moedas nacionais

Casa da Moeda

estabelecimento oficial onde se fabrica moeda

pagar na mesma moeda *figurado*

retribuir do mesmo modo

(Do latim *monēta-*, «moeda; dinheiro cunhado»)

Figure 4.8: Infopédia's definition of the word *moedas*.

Figure 4.8 shows the result of using Infopédia to obtain information about the word *moedas* (coins). Almost all the definitions provided by this dictionary are presented in similar fashion, with three distinct elements:

- A modifier stating the tense, gender and/or number of the word;
- A subset of definitions;
- A list of compounds, collocations and idiomatic expressions containing the word.

[6] `http://www.infopedia.pt` (last visited in August 2013)
[7] `http://www.priberam.pt/dlpo/` (last visited in August 2013)

Since we are only interested in keeping the first two elements, the list of idiomatic expressions is discarded. This process is done by extracting the relevant text from the web-page and delimiting each definition. After this, a regular expression is used to identify the first definition where the word appears. This definition typically corresponds to the first idiomatic expression, from which the remainder of the text can be discarded. Even if that is not the case, the definition is still discarded, because a clue containing the answer is not suitable for the puzzle at hand.

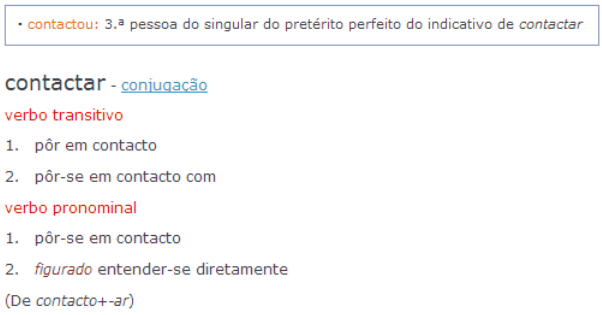

Figure 4.9: Infopédia's definition of the word *contactou*.

Conjugated verbs, however, pose another problem. For example, Figure 4.9 shows the result of using Infopédia to obtain the definition of the word *contactou* (contacted). Since there are no idiomatic expressions, all the definitions presented may be kept. However, the top line, which refers to the verb form, clearly states which verb is being conjugated, which makes the remaining definitions useless, since the information in the modifier is sufficient to identify the word. Once more, using regular expressions, the non-essential information in this component is discarded.

This trimming process, which is done in order to eliminate any unnecessary or redundant information, differs from word to word. This required each word do be analysed by hand. For example, in the modifier of the word *moedas*, the information regarding the gender and the part of speech is not particularly helpful because on this specific case, it has no impact on how the word is transformed given its base form. The number however, indicates that some changes need to be made, namely, the addition of the *"-s"* suffix. Furthermore, the existence of information that states which word undergoes the modification (*"(...)de moeda"*), defeats the purpose of trying to find the target word using the definitions, since it clearly indicates what the answer is.

The results of the formatting operations for the words *moedas* and *contactou* can be observed in Figures 4.10 and 4.11. The percent sign is used to indicate whether a modifier is present or not, and the semicolon character is used to separate the existing definitions.

The result of this whole process was then stored in the same directory as the exercise. This was done to speed up the lookup process.

```
moedas-Plural.%
Peça geralmente metálica, cunhada com autorização legal,
que serve para realizar transações financeiras.;
Unidade monetária em vigor em determinada região.;
Tudo o que representa um valor pecuniário.;
ECONOMIA instrumento que exerce as funções de meio de pagamento,
medida de valor e meio de reserva.;
ECONOMIA conjunto dos meios de pagamento de uma economia.
```

Figure 4.10: Formatted definition of the word *moedas*.

```
contactou-3a pessoa do singular do pretérito perfeito do indicativo.%
Pôr em contacto.;
Pôr-se em contacto com.;
Pôr-se em contacto.;
FIGURADO entender-se diretamente.
```

Figure 4.11: Formatted definition of the word *contactou*.

Since these resources are already available, the second difficulty parameter is taken into account in order to select the most appropriate definition. If the user chooses the easy difficulty level, the definition of the given word will be the result of concatenating the modifier (in case it exists) and the first sentence ending with ".;", that is, the first definition. For example, the easy level definition for the word *moedas*, after the auxiliary characters are removed, would be *"Plural. Peça geralmente metálica, cunhada com autorização legal, que serve para realizar transações financeiras."* (Plural. Typically metallic item, coined with legal authorization, used to conduct financial transactions.).

If the user chooses the medium level, the second part of the definition is chosen randomly from all the available definitions, excluding the first and the last one. Finally, if the hard level is chosen, the last definition is appended to the modifier. If only one definition is available, it is chosen regardless of the level. If there are two, both medium and hard levels use the second one.

4.2.4 Support

At any point during the process of solving the puzzle, the user has access to three different help functionalities. These are accessible through a series of buttons, as can be seen in Figure 4.5.

Colorir (Color)

The first of these is available via the *Colorir* (Color) button, which changes the grid's default color scheme, painting each cell's border green or red depending on whether its contents (each cell's specific syllable) are correct or wrong.

1					con	vo	ca	ção
2					su	mári		o
3					me	li	tar	
4	fi	nã	ci	a	dor			

Figure 4.12: Example of the *Colorir* help being used.

An example of this operation can be found in Figure 4.12

Dica (Hint)

The second help option called *Dica* (Hint), allows the user to automatically obtain the contents (the correct answer) of a group of cells. This is done by filling the target cells with an 'X' character. If no cells are filled in this way, an error message appears in order to remind the user on how to use this feature.

An example of this feature being used on the first and the fourth cells fourth horizontal word, can be found in Figure 4.13

1	e			dos	
2	di			í	
3	to		X	a	
4	ri	X			te
5	al			do	

1	e			dos	
2	di			í	
3	to		vi	a	
4	ri	gi			te
5	al			do	

Figure 4.13: Example of the *Dica* help feature being used on the third cell of the third horizontal word and on the second cell of the fourth horizontal word.

Mais Definições (More Definitions)

The final help feature called *Mais Definições* (More Definitions), is implemented in a similar fashion to the *Dica* assistance. Users are once again required to fill cells with an 'X' character but this time, instead of filling the cells with the correct answer, the complete definitions, for the words to which the cells belong to, are presented in the blackboard. The original definitions are removed from the blackboard and its counterparts are appended to the original definitions. This is done to highlight these new definitions, in contrast to the remaining ones.

Once again, an example of this operation can be found in Figure 4.14

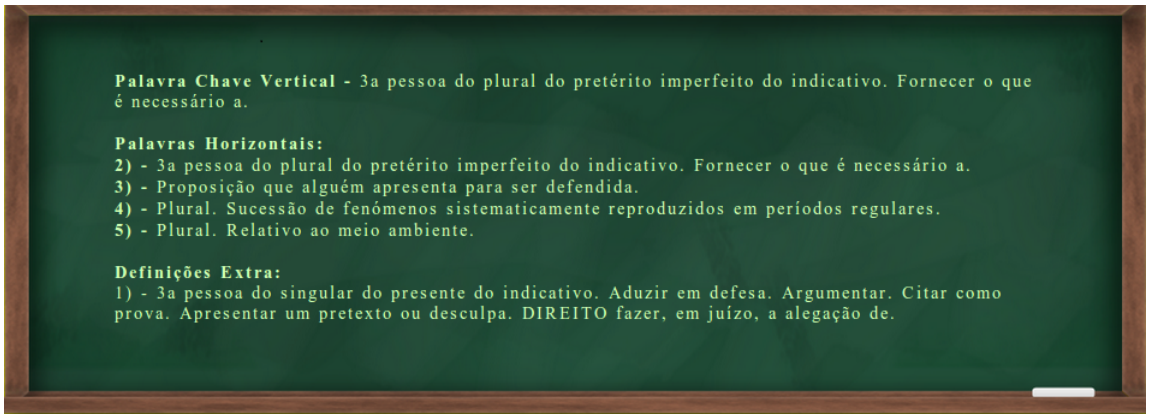

Figure 4.14: Example of the *Mais Definições* help feature being used on the first horizontal word.

4.2.5 Score and Feedback

After submitting the puzzle grid, by pressing the button "*Continuar*" (continue), the user is presented with a feedback screen, which includes both the submitted and the solution grids. An example of this screen can be seen on Figure 4.15 The fact that they are both displayed side by side and highlighted accordingly, allows for an easier and more practical comparison between what the user's answer was and, the solution to the puzzle. Information on the number of hints used, as well as the number of both correct and incorrect answers is also made available.

When combined with the difficulty levels selected in the beginning of the exercise, these parameters are used to compute the puzzle's maximum score and the user's score. To discourage an excessive use of the help features, each hint has a negative impact on the final score.

Each puzzle's maximum score is calculated according to Expression 4.1. The user's score is computed according to Expression 4.2. By clicking on the *Pontuação Detalhada* (Detailed Score) button, the user has access to a table which contains these formulas and the values corresponding to each variable (as well as the results).

$$Maximum\ Score\ =\ \#Empty\ Cells * (Definition\ Level + Grid\ Level) * 5 \tag{4.1}$$

$$Score\ =\ (\#Correct\ Cells - \#Hints) * (Definition\ Level + Grid\ Level) * 5 \tag{4.2}$$

Besides this, the user also has the option to begin a new exercise, return to the student's menu, or to fill-out the usability and satisfaction questionnaire used on this project's evaluation stage.

Figure 4.15: Crossword puzzle exercise feedback screen.

4.3 Derivation Exercise

Derivation (id. in English) is one of the processes languages use to form new words from other existing words or from root morphemes (stems). In this view of derivation, a base form undergoes a morphological process that adds a morpheme yielding a new, derived word. In Portuguese, these morphemes are affixes, prefixes, placed before the base, and suffixes, attached to the end of the base. Suffixes often change the part-of-speech of the base word.

In this exercise, the main focus is to identify words formed through prefixal derivation (that is, derivation by prefix addition to a base form) that constitute exceptions to the general rules of the prefixation process. These exceptions occur typically with certain prefixes such as, but not limited to, *a(n)–*, *de(s)–*, *di(s)–* and *i(m)–*. These prefixes often change the meaning of the base word regularly, yielding its antonym or opposite sense. However, prefixes like *a(n)–* can also express a privative meaning (lack of a quality), e.g. *acrítico* (acritical).

A common example can be found in the word *correto* (correct). This word has its meaning negated if we apply the prefix *in–* changing it into *incorreto* (incorrect). However, there are certain exceptions to this rule and the application (or presence) of negative prefixes in certain words, or stems, will not always mean that the given word has its meaning negated if said prefix is removed. For example, the word *infiltrado* (infiltrated), does not mean the opposite of *filtrado* (filtered).

The welcome web-page can be found in Figure 4.16 and an example of the exercise can be found in Figure 4.17.

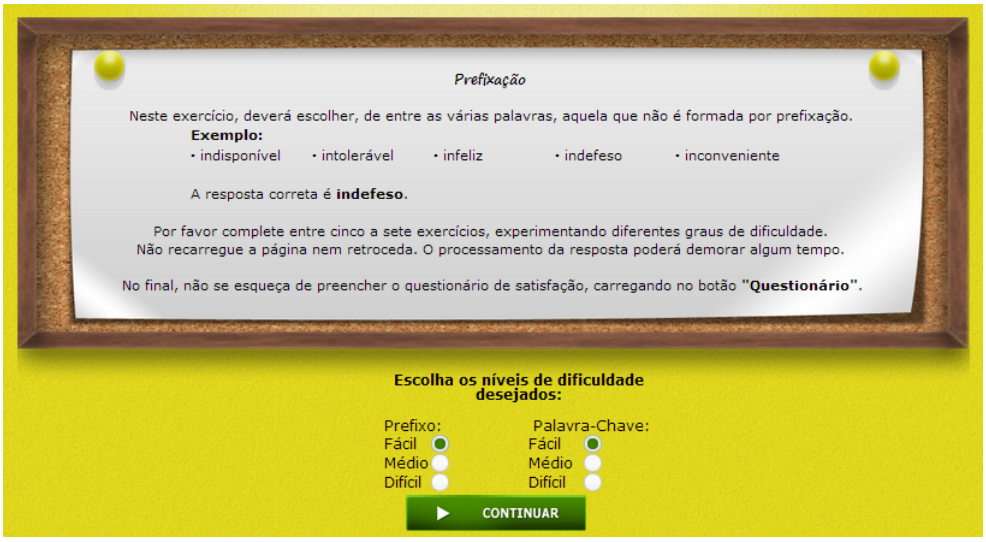

Figure 4.16: Derivation exercise welcome web-page.

Similar to what occurs in the syllable division process, derivation by prefix addition is an operation which depends on a series of rules that specify which prefixes can be applied to certain words. However, since the goal is to identify the exceptions to these rules, namely the words seemingly formed by prefix addition, a pattern cannot be applied to easily identify these exceptions.

With that being the case, both the exceptions and the distractors used in this exercise were the result of the update process of the LexMan (Vicente, 2013) dictionaries. LexMan is part of STRING (Mamede et al., 2012), the natural language processing chain of L^2F. Whenever a text is processed by STRING, the LexMan module is responsible for the tokenization and morphological analysis steps. The latter process is done using transducers, and it is also during this process that prefixed words are identified. Basically, each time a word is morphologically analysed, in order for the corresponding morpho-syntactical (part-of-speech) tag to be assigned, the dictionary transducer also "checks" whether the beginning of the word corresponds to a prefix.

This process can be observed in Figure 4.18, where the paths from state 1 to state 20 correspond to the identification of a prefix. The empty transition from state 0 to state 20 (*epsilon*) corresponds to the opposite.

Regardless of whether or not a prefix was identified, the remainder of the word is parsed and a tag is assigned. The numbers in some transitions such as the one from state 45 to state 46, represent the indexes which are later used to locate the words in the corresponding grammatical category files.

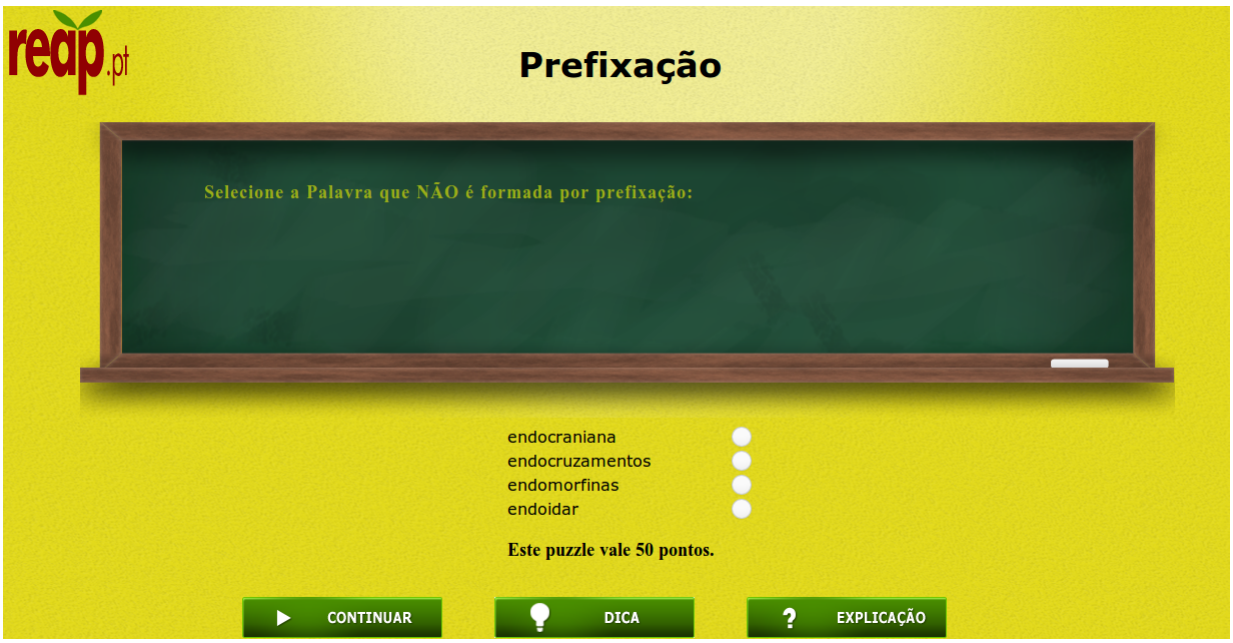

Figure 4.17: Derivation exercise interface.

The list of valid combinations of prefixes and words was manually created by hand, in an effort to remove variations of certain lemmas from the dictionary transducer and to just keep in it other words that are not formed by prefix addition. For example, the word *reescrever* (rewrite) is a variation of the lemma *escrever* (to write), because it is the result of applying the prefix *re–* to said lemma. Since it is redundant to have both *reescrever* and *escrever* in the dictionary transducer, with the system now able to identify these words, the variations in question were removed from the dictionary list. This process is explained in greater detail in (Vicente, 2013).

For this exercise, the words which were valid combinations of prefixes and lemmas were used as the distractors. Those which were not, were used as the exceptions (target words).

Figure 4.19 shows the steps that are taken in order to generate each puzzle.

4.3.1 Prefix Selection

The process of updating the LexMan dictionaries yielded 522 distinct files containing exceptions, with each file corresponding to a different prefix. On the opposite end, 15328 distractors, were also made available. Given the fact that from this list of prefixes, some only contained a handful of exceptions and/or the fact that others had very few "assignable" distractors, this list of 522 files was trimmed. Furthermore, it was necessary to classify each prefix according to a difficulty level.

In order to reduce the number of prefixes used in this exercise, the list of files was filtered to exclude

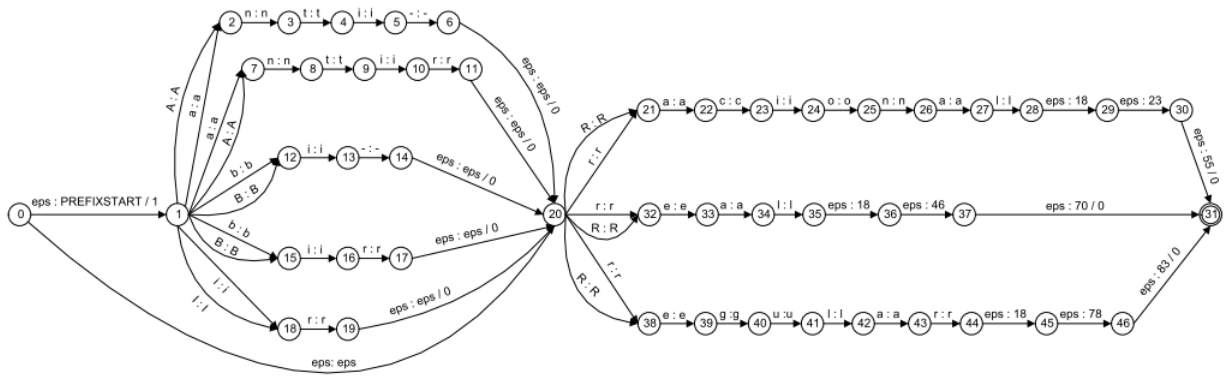

Figure 4.18: Part of the LexMan Transducer responsible for identifying adjectives beginning with the letter *r*.

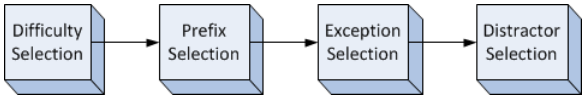

Figure 4.19: Derivation exercise generation steps.

those which contained less than ten exceptions and less than ten distractors, with the result being a list of fifty two files. Besides this, the difficulty levels were assigned according to the pseudo-code found in Listing 4.1.

Listing 4.1: Prefix difficulty level attribution formula.

```
IF (Distractors>100 AND Exceptions>100) LEVEL = Easy;
ELSE IF (Distractors < 50 OR Exceptions < 20) LEVEL = Hard;
ELSE LEVEL = Medium;
```

The complete list of prefixes, including the number of exceptions, distractors and difficulty level can be found in Annex B. A condensed view of these values according to difficulty level can be seen in Table 4.4

Difficulty Level	Distractors	Exceptions
Easy	6892	4370
Medium	4498	1051
Hard	1607	1202

Table 4.4: Total number of distractors and exceptions for each difficulty level, in the Derivation exercise.

Finally, with this being said, the first stage when generating a puzzle consists of randomly choosing

a prefix from the list, depending on the difficulty level selected by the user. Similar to what happens in the Crossword Puzzle Exercise, the user is asked to choose between three difficulty levels for two different parameters: the prefix and the exception (target word). Once more, this is done in the index page of the exercise.

4.3.2 Exception Selection

Further parametrization was required to distinguish between easy exceptions and the more difficult ones. This was done by calculating the frequencies of each exception (uni-gram count), using the total number of forms found in the Linguateca[8] corpora[9]. Among these are CETEMPúblico (Santos & Rocha, 2001), DiaCLAV[10], which corresponds to a compilation of online articles extracted from several regional Portuguese newspapers, and Vercial, a corpus consisting of 309 literary works from 55 Portuguese authors, scanned and made available by the Vercial Project[11]. Since these files are updated frequently, the file[12] corresponding to the totals is also updated. For this reason, the 28 of March 2013 version was stored locally and used for the entirety of this process. Words which did not appear in this list were assigned a count value of one.

Listing 4.2: Exception difficulty level attribution formula.

```
1 IF (Occurrences > 1000) LEVEL = Easy;
2 ELSE IF (100 < Occurrences <= 1000) LEVEL = Medium;
3 ELSE LEVEL = Hard;
```

Difficulty level assignment was done in similar fashion to what happened with the prefixes. Words are categorized according to the pseudo-code formula from Listing 4.2. This is where the second difficulty parameter comes in play. Depending on the user's choice in the index screen, an exception with the matching difficulty level is randomly extracted from the list of exceptions corresponding to the prefix chosen in the previous stage.

4.3.3 Distractor Selection

In multiple-choice questions such as this one, the process of selecting the distractors is typically done according to the nature of the problem itself, and the correct answer. In this specific exercise, three different distractors are chosen for each puzzle. The first one, which can be referred to as the "easy distractor", is, once again, randomly extracted from the distractor file. This file also contains, for each

[8] http://www.linguateca.pt/ (last visited in August 2013)
[9] http://www.linguateca.pt/acesso/contabilizacao.php (last visited in August 2013)
[10] http://www.linguateca.pt/acesso/corpus.php?corpus=DIACLAV (last visited in August 2013)
[11] http://alfarrabio.di.uminho.pt/vercial/ (last visited in August 2013)
[12] http://www.linguateca.pt/acesso/tokens/formas.total.txt (last visited in August 2013)

distractor, the corresponding count in the total forms file. For this distractor specifically, only words with a count greater than 62 are considered. Several other values for this parameter were tested, however, since there was a big dispersion in the number of occurrences, the average number (62,221) was chosen as the threshold value. The second distractor, labeled as the "difficult distractor", is chosen in a manner similar to its "easy" counterpart. With that being said, only words with an occurrence count of 1 are eligible, that is, words that only appeared once, or did not appear at all in the texts of the corpus, are suitable to be selected.

Finally, and contrary to what happens in the previous cases, the third and final distractor is chosen according to the word itself and not its occurrence count. Once again, the Levenshtein Distance, specifically the same PHP implementation[13] as the one used in the Crossword Puzzle Exercise, was chosen as a similarity standard. The initial threshold is set to the length of the exception. After this, ten distractors are randomly chosen from the distractor list, and the one with the lowest Levenshtein Distance when compared to the threshold value, is chosen as the "similar" distractor. If there are none that pass these conditions, that is, if none of them have a value lower than the length, of the target word, one of the ten distractors is randomly chosen.

It is important to note that both processes of difficulty level attribution are based solely on quantitative data and therefore, do not always guarantee that a given word with a lower number of occurrences when compared to another, is indeed more difficult than said word. Since no human evaluation was done in order to categorize the prefixes and the distractors, other parameters such as the nature of the prefix, *monossémico* (only one meaning) or *polissémico* (several distinct meanings), the application of the hyphen character and the need to adjust the base form in order to apply the prefix, are not considered. However, given the fact that these operations are done according to a set of values and parameters, it allows for tweaks, in the long term, if deemed necessary.

After the end of this step the distractors, along with the correct answer, are shuffled before being presented to the user.

4.3.4 Support

During the process of solving the exercise, the user has the option of removing a random distractor, by pressing the *"Dica"* (Hint) button. This can be done up to three times, leaving the final correct answer, as the remaining option. This operation has a direct impact on the final score of the exercise. This is done in order to discourage the excessive use of this feature.

Besides this, accessing the help screen, by clicking on the *"explicação"* (explanation) button, allows

[13] `http://php.net/manual/en/function.levenshtein.php` (last visited in August 2013)

to user the view information on the prefix being used in the exercise. This information was extracted from the Priberam[14] on-line dictionary, and stored locally in order to speed up the lookup process.

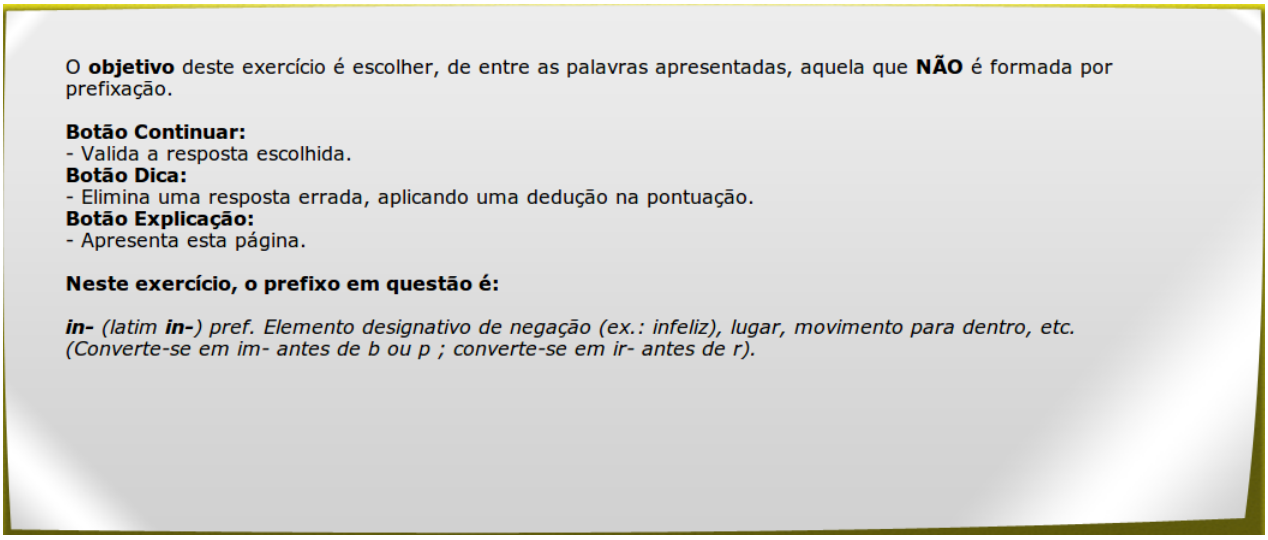

Figure 4.20: Derivation puzzle help screen.

Figure 4.20 shows an example of an help screen, containing information on the prefix .

4.3.5 Score and Feedback

After choosing one of the available options, the user is presented with a different feedback screen, depending on whether the correct answer was chosen or not.

If that is indeed the case, a screen similar to the one found in Figure 4.21 is shown.

However, if the user chooses an incorrect answer, a screen similar to the one found in Figure 4.22 is shown. In either case and whenever they are available, real sentences using the correct and/or incorrect words are presented. These are extracted from the CETEMPúblico corpus (Santos & Rocha, 2001), using the Ensinador[15] Tool from Linguateca. Since the use of this tool requires a form to be filled each time a query is made, the Snoopy[16] PHP Class was used to process these forms and to extract the relevant text from the output web-page.

$$Maximum\ Score\ =\ (Definition\ Level + Grid\ Level) * 5 \tag{4.3}$$

[14] `http://www.priberam.pt/dlpo/` (last visited in August 2013)
[15] `http://www.linguateca.pt/Ensinador/index.pl?corpus=CETEMPUBLICO` (last visited in August 2013)
[16] `http://sourceforge.net/projects/snoopy/` (last visited in July 2013)

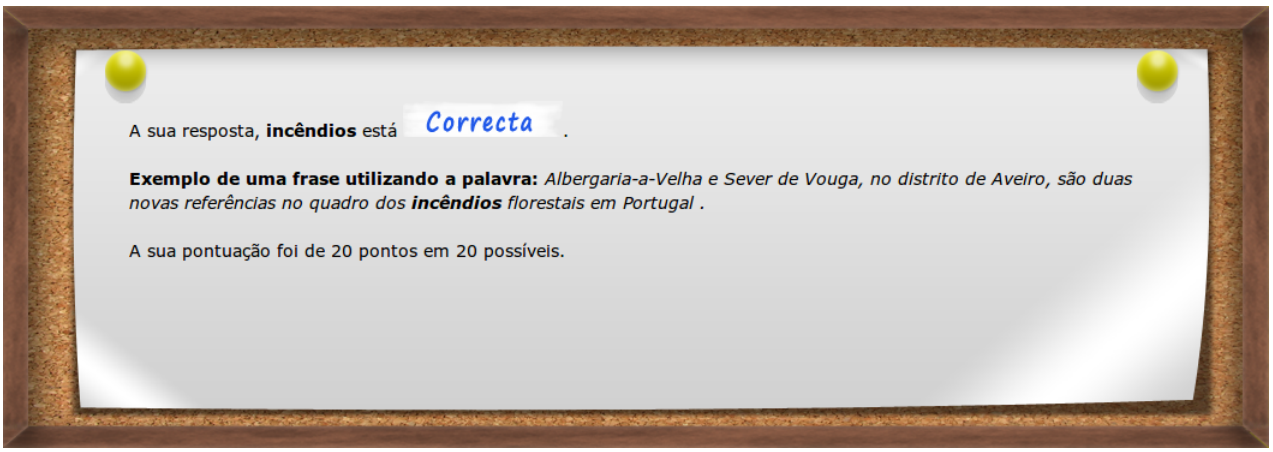

Figure 4.21: Derivation puzzle feedback screen for the correct answer.

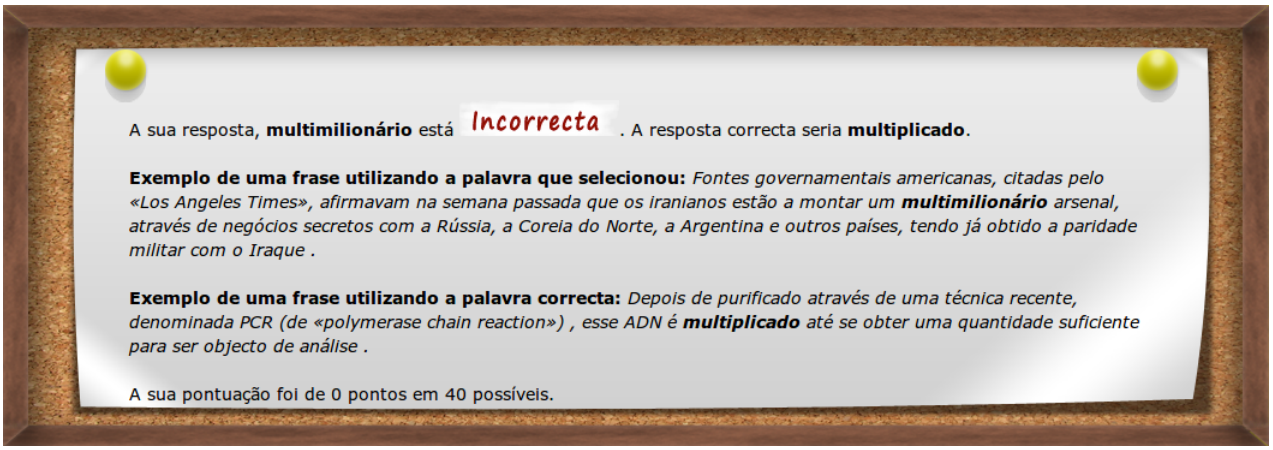

Figure 4.22: Derivation puzzle feedback screen for an incorrect answer.

Listing 4.3: Derivation Exercise Final Score Algorithm.

```
IF (Answer == CorrectAnswer)
{
    IF (UsedHints == 0) Score = MaxScore;
    ELSE IF (UsedHints == 1) Score = MaxScore/2;
    ELSE IF (UsedHints == 2) Score = MaxScore/4;
    ELSE Score = 0;
}
ELSE Score = 0;
```

Besides this, the user is also presented with his score and the maximum score of the puzzle he or she just solved. Each puzzle's maximum score is calculated according to Expression 4.3. The user's score is computed according to the pseudo-code in Listing 4.3. Users are penalized not only for choosing an incorrect option, but also for using the help feature.

5 Evaluation

In this chapter, the evaluation process is described and its results are presented.

5.1 Evaluation Setup

During the project's initial stage, a more complex evaluation process was planned. Firstly, users (test subjects) would have been asked to evaluate several components of the exercises. This would have been done by applying a Likert scale (Likert, 1932) to these components, which would be later summed in order to obtain results on the overall quality of the exercises. Said components included the quality of the provided clues, the correctness and adequacy of the exercises and the time each exercise required for completion, for example. Secondly, the users would be divided into two groups. This operation would allow a comparison to be drawn, between students which would only solve automatically generated exercises by the REAP.PT system and those who would also used other resources such as Portuguese textbooks. The objective would be to assess how the REAP.PT system fared against the more traditional resources, when used as a learning tool. Finally, the system's usability would also be evaluated via a survey based on the *USE Questionnaire* (Lund, 2001). However, due to time and human constraints, while the first stage had to be modified, the second stage of this evaluation process had to be dropped altogether and, finally, the third stage was expanded.

Two different types of evaluation were carried out for both exercises, crowd-source testing and a questionnaire.

5.1.1 Crowd-sourced Testing

In order to obtain a significant number of results, websites were made available for testing both exercises. The original plan was to have both native and non-native Portuguese users solve several exercises. Depending on the results obtained, this process would help determine whether the developed exercises were deemed adequate for foreign students, or if they were too difficult for them to solve. More specifically the goal was to conclude whether the differences between the difficulty levels, would also apply to the non-native speakers, allowing them to correctly solve part of the generated exercises. However, even tough some attempts were made, it was not possible to find non-native Portuguese speakers to

carry out the tests at the time of this document's elaboration.

Users were asked to solve between three and five puzzles for the Crossword exercise and between five and seven puzzles for the Derivation exercise, in order to try out different combinations of difficulty levels. However, it is important to note that no limitations were made to prevent the users from solving too many or too few exercises, that is why the number of solved puzzles exceeds the value one might have expected by multiplying the number of answers to the questionnaires times the upper limit of the recommended exercises to solve. In similar fashion, no restrictions were applied in order to force the users to try out different combinations of difficulty levels. Again, this was done in order to allow them to progress and to solve more complex puzzles at their one pace.

Several parameters were recorded for each solved exercise. Furthermore, due to the architecture of both puzzles it was possible to record approximately the same information. These parameters consist of: Score, maximum (possible) score, difficulty levels, the number of hints that were used and, in the derivation exercise, the prefix used.

5.1.2 Questionnaire

In order to evaluate other aspects of the exercises such as their interface, which could not be assessed simply by looking at the raw data extracted from the solved puzzles as it requires specific input from the users regarding what they like and dislike about the experience, questionnaires were introduced after the feedback screens of both exercises, if the users followed the corresponding links. These questionnaires were a combination of the *USE Questionnaire* (Lund, 2001) which stands for Usability, Satisfaction and Ease of use, and on GameFlow (Sweetser & Wyeth, 2005), a model used for evaluating the player's enjoyment in games. Even tough GameFlow is commonly applied to more complex games, such as the ones with time constraints and those which require the user to perform several different tasks, it was possible to apply several of its aspects to the developed exercises.

The majority of the questions follow a five-point Likert rating scale (a psychometric scale). With this scale, users are asked to rate their agreement with a certain sentence or question, ranging from *strongly agreeing with it* to *strongly disagreeing with it*.

Both questionnaires were comprised of a set of multiple choice questions and a free-form text box. The latter allowed for personalized feedback.

Question	Scale	Crosswords	Derivation
Qual a sua idade? (What is your age?)	Any positive integer	✔	✔
O português é a sua língua materna? (Is Portuguese your mother tongue?)	Yes/No	✔	✔
Qual é o seu grau de escolaridade? (What is your educational level?)	Multiple-choice	✔	✔
O sistema é fácil de utilizar. (The system is easy to use.)	Likert	✔	✔
Percebi rapidamente o objectivo. (I understood the goal quickly.)	Likert	✔	✔
Depois de identificar a palavra, tive dificuldades em dividi-la. (After identifying the word, I had trouble dividing it.)	Likert	✔	
Os exercícios são demasiado fáceis. (The exercises are too easy.)	Likert	✔	✔
O feedback (retorno) dado pelo sistema é suficiente. (The feedback provided by the system is sufficient.)	Likert	✔	
Notei diferenças entre os graus de dificuldade. (I noticed there were differences across the difficulty levels.)	Likert	✔	✔
Aprendi alguma coisa ao utilizar o sistema. (I learnt something using the system.)	Likert	✔	✔
Diverti-me ao utilizar o sistema. (I had fun using the system.)	Likert	✔	✔
Qual a ajuda mais útil? (Which "helping" feature was the most useful?)	Multiple-choice	✔	
Qual a ajuda menos útil? (Which "helping" feature was the least useful?)	Multiple-choice	✔	
Apreciação global do sistema. (Global system appreciation.)	Multiple-choice	✔	✔

Table 5.1: Questionnaire composition.

The contents of the questionnaires can be seen in Table 5.1. The elements based on the USE factors and the GameFlow items are also displayed in Table 5.2. The questionnaires web-pages can be found in Annexes D and E.

Question	*USE* factor	GameFlow element
O sistema é fácil de utilizar. (The system is easy to use.)	Ease of use	
Percebi rapidamente o objectivo. (I understood the goal quickly.)		Clear goals
Os exercícios são demasiado fáceis. (The exercises are too easy.)		Challenge
O feedback (retorno) dado pelo sistema é suficiente. (The feedback provided by the system is sufficient.)		Feedback
Notei diferenças entre os graus de dificuldade. (I noticed there were differences across the difficulty levels.)		Challenge
Aprendi alguma coisa ao utilizar o sistema. (I learnt something using the system.)	Usefulness	Player skills
Diverti-me ao utilizar o sistema. (I had fun using the system.)	Satisfaction	
Apreciação global do sistema. (Global system appreciation.)	Satisfaction	

Table 5.2: Questionnaire elements.

5.2 Crossword Puzzle Exercise Evaluation

5.2.1 Crowd-sourced Test Results

During the exercise evaluation stage, a total of 407 puzzles were submitted. Since the system only recorded puzzles which were either correctly or incorrectly solved, this number does not include puzzles that were started but never finished. This was done because during the evaluation stage, users were encouraged to always submit their puzzles, regardless of their performance.

		Definitions			
		Easy	Medium	Hard	Total
Grid	Easy	208	15	10	233
	Medium	33	56	2	91
	Hard	28	2	53	83
	Total	269	73	65	407

Table 5.3: Total number of submitted crossword puzzles by difficulty level.

The total number of submitted puzzles, according to the selected difficulty levels, can be seen in Table 5.3. The easiest possible combination was the one pre-selected by default in the index page. This fact explains the high number of puzzles with these parameters. As it was previously mentioned, no restrictions were implemented in order to prevent users from always solving puzzles of the same combination of difficulty levels. This fact explains in part why this specific combination of difficulty levels has an higher value than its counterparts (51.1% of the total number of submitted puzzles). Apart from this,

users also preferred to solve puzzles with the remaining equivalent difficulty levels, namely medium-medium (13,8%) and hard-hard (13,0%), instead of mixing and matching from the available possible combinations.

		Definitions		
		Easy	Medium	Hard
Grid	Easy	65.13%	53.19%	33.33%
	Medium	68.09%	47.41%	81.25%
	Hard	65.25%	72.97%	33.37%

Table 5.4: Average score of the submitted crossword puzzles, according to the difficulty levels.

Table 5.4 shows the score ratio for the total number of puzzles submitted in each possible combination. This value corresponds to the score the user achieved, over the maximum possible score in that given puzzle. If we only take into account the combinations with the highest number of submitted puzzles (easy-easy, medium-medium and hard-hard), one can clearly see that as the puzzles become more difficult, the values in Table 5.4 decrease, indicating that the combination of more difficult parameters generate more difficult puzzles. Furthermore, the results obtained for the combinations medium-easy and hard-easy are not that different. This fact may suggest that the definition difficulty level has an higher impact on the overall difficulty of the puzzle than the grid difficulty level.

		Definitions		
		Easy	Medium	Hard
Grid	Easy	1.43	1.8	1.1
	Medium	2.3	3.77	1.5
	Hard	1.46	0.5	5.4

Table 5.5: Average number of used hints for the submitted crossword puzzles, according to the difficulty levels.

The average number of used hints per difficulty level can be found in Table 5.5. Similar to what happens in Table 5.4, if we only take into account the most frequently used combinations, the number of used hints increases as the puzzles get more difficult. This is also a good indicator that the differences between the difficulty levels are notorious, since users felt the need to use the help features more frequently as the puzzles got more difficult.

5.2.2 Questionnaire Results

The questionnaire results for the crossword puzzle exercise, can be found in Figures 5.1 to 5.11. In total, 59 users filled-in the form. Their average age was 23.1, ranging from 18 to 41 years old. 73% of the

users had completed college education, with the remaining 27% having graduated high school. Results for the third question, *"O português é a sua língua materna?"* (Is Portuguese your mother tongue?), were omitted because all inquired users were Portuguese native speakers.

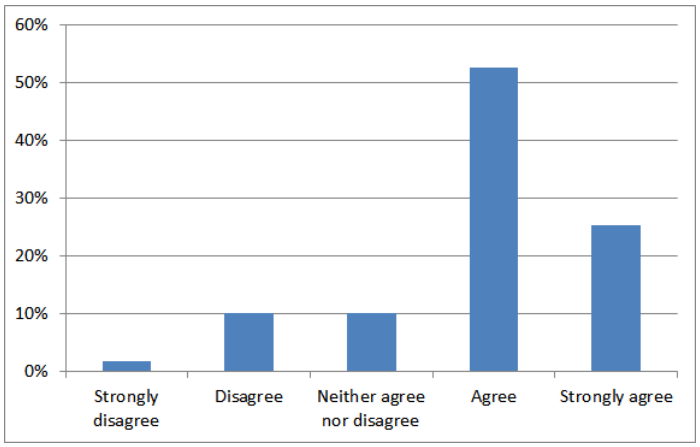

Figure 5.1: Crossword Questionnaire results for the statement "The system is easy to use.".

Regarding the statement "The system is easy to use", 78% of the users agreed with it, with only 12% disagreeing with it.

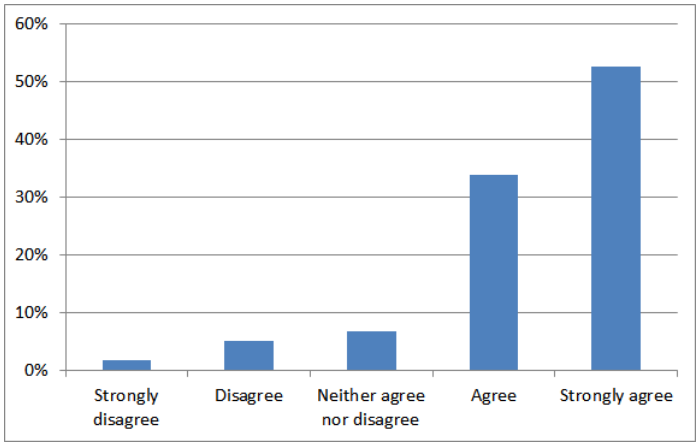

Figure 5.2: Crossword Questionnaire results for the statement "I understood the goal quickly.".

Higher percentages were achieved in the questions relating to the objectives of the exercise. 86% of the users declared to have quickly understood the goal of the exercise (Figure 5.2), and 93% stated that they did not have any problems dividing the word. In fact, out of 59 inquired users, only one of them claimed the opposite, as can be seen in Figure 5.3.

Most users (over 80%) either disagreed or chose not to support the statement which claimed that the exercises were too easy (Figure 5.4). This percentage validates the approach used to generate challenging

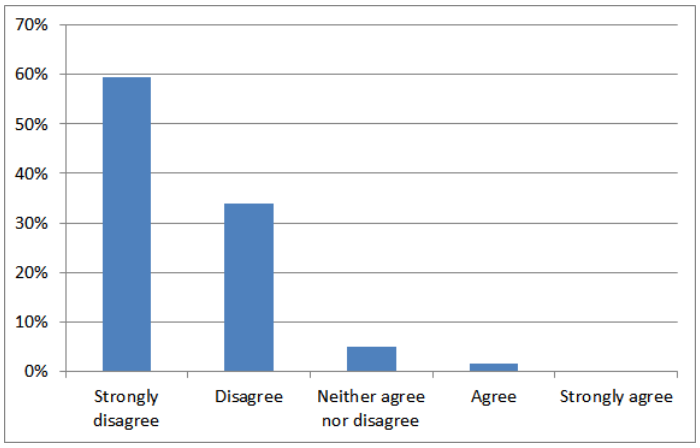

Figure 5.3: Crossword Questionnaire results for the statement "After identifying the word, I had trouble dividing it".

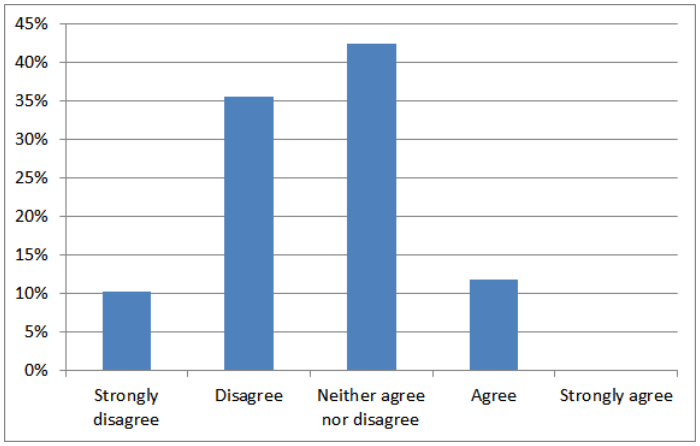

Figure 5.4: Crossword Questionnaire results for the statement "The exercises are too easy".

puzzles regardless of the combination of the difficulty levels. Results regarding the scalability of these can be seen in Figure 5.6, in which 80% of the users claim to have noticed differences across the different possible combinations.

The feedback provided by the system also proved to be adequate, with 68% of the users at least agreeing with this statement.

Figures 5.7 and 5.8 concern the usefulness and satisfaction factors respectively. Once again, positive scores were achieved in the corresponding question. Only 24% of the users claimed to not have learned anything by using the system. Even tough this value is not particularly high, it is important to note that similar to what happens in traditional crosswords puzzles, the act of guessing words given a group of clues and dividing them afterwards, are tasks which typically become easier when they are carried

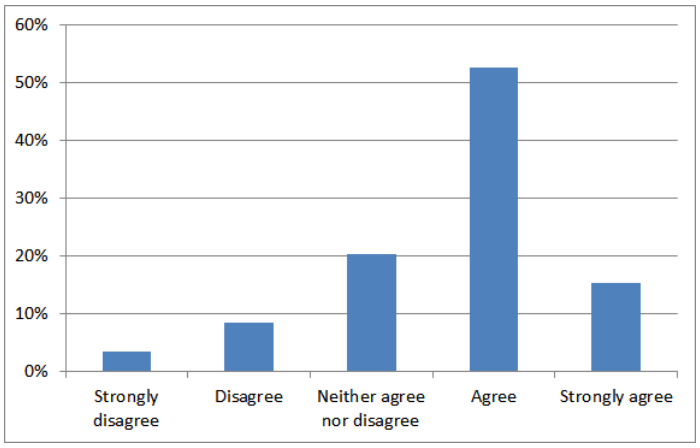

Figure 5.5: Crossword Questionnaire results for the statement "The feedback provided by the system is sufficient".

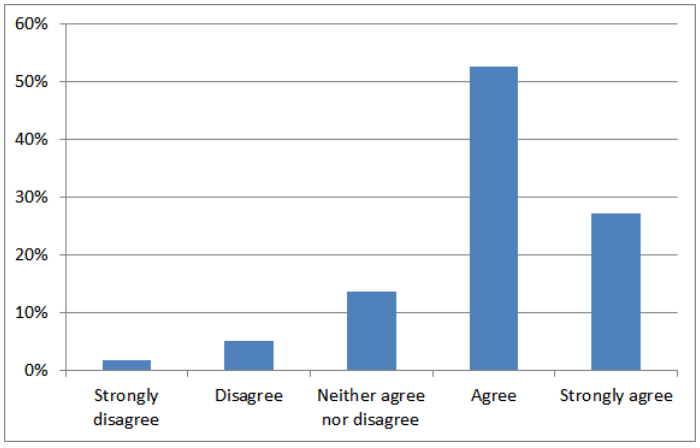

Figure 5.6: Crossword Questionnaire results for the statement "I noticed there were differences across the difficulty levels.".

out frequently. It is possible that a lower value would have been achieved if users were asked to solve puzzles, using the system, over a longer period of time or with an higher number of recommended puzzles to solve. Regarding the satisfaction of the users, only 15% claimed not to have fun while using the system, while 58% stated the opposite. The overall score is positive.

Concerning the help features, the hint option was the preferred one, with 53% of the users considering it to be the most useful (the highest value in Figure 5.9) and only 15% of the inquired population considering it the least useful (the lowest value in Figure 5.10).

Finally, regarding the overall system appreciation, 71% of the users claimed to be either satisfied or very satisfied the system. The fact that only 3 users, corresponding to 5% of the inquired population, claimed to be somewhat dissatisfied with it and the fact that not a single user claimed to be very

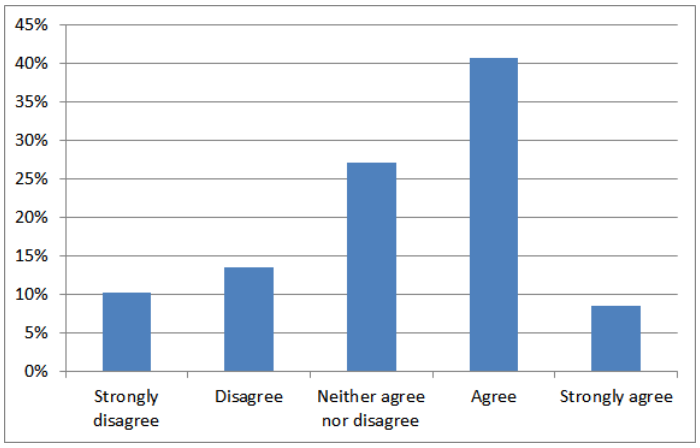

Figure 5.7: Crossword Questionnaire results for the statement "I learnt something using the system".

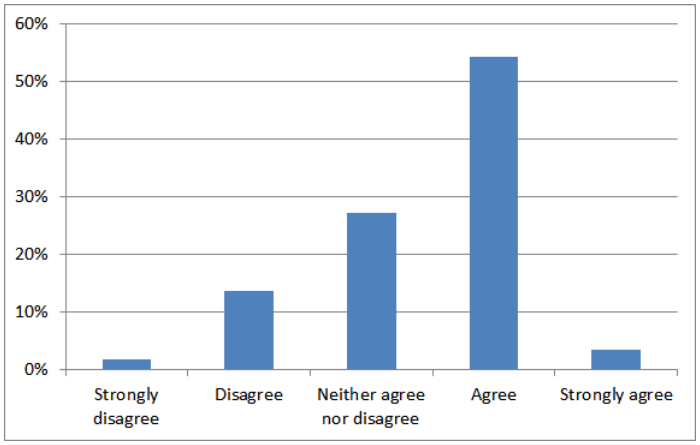

Figure 5.8: Crossword Questionnaire results for the statement "I had fun using the system".

dissatisfied with the system, validates the approach followed in this project.

5.2.3 Questionnaire Comments

In the final question of the questionnaire, users were asked to suggest modifications to the exercise and to identify what they would like to see improved. Several issues were raised and some suggestions were made.

Difficulty levels

Several users noted that sometimes, the medium difficulty level definitions were easier than the easy level ones. Even tough this may sometimes occur due to the discrepancies on the difficulty of the words

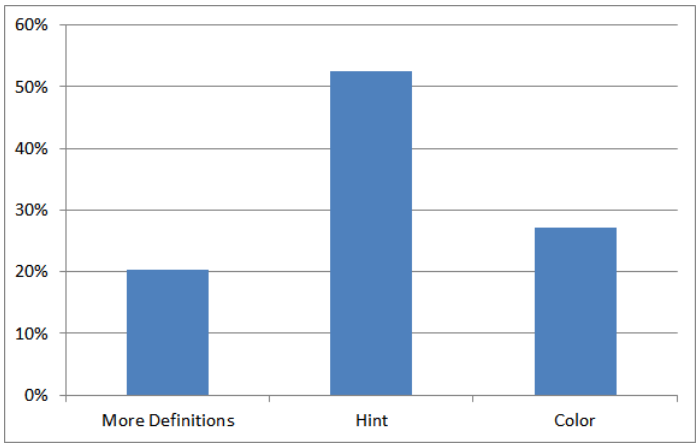

Figure 5.9: Crossword Questionnaire results for the question "Which "helping" feature was the most useful?".

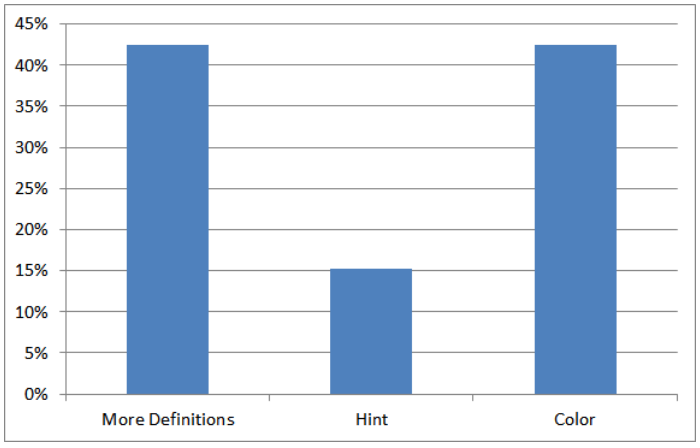

Figure 5.10: Crossword Questionnaire results for the question "Which "helping" feature was the least useful?".

themselves, these complains can be attributed to the way in which all definitions are ordered. The *Infopédia* dictionary considers definitions specific to a certain context to be more difficult than other general and/or more "generic" definitions. Due to that, it is possible for users accustomed to deal with specific contexts to be more familiar with said definitions.

Some users also complained about the fact that the definitions from the "hard" level were too difficult. While our objective was not to generate puzzles which would be impossible to solve, these concerns validate the answers to the statement "I noticed there were differences across the difficulty levels." of the questionnaire (Figure 5.6).

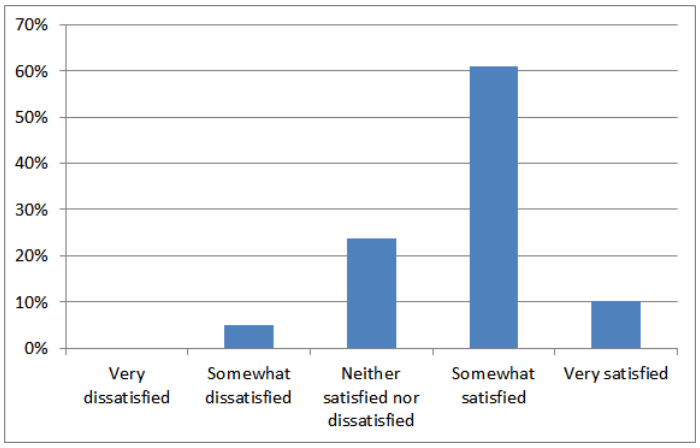

Figure 5.11: Crossword Questionnaire results for the statement "Global system appreciation.".

Interface modifications

Some users had problems with the interface. Their main concern was the lack of information presented before the puzzle was generated, namely, the purpose of all the buttons in the exercise page and how to use the functionalities they provided. Some of them also noted that the support buttons should have been placed elsewhere, in order to prevent confusion, and/or redesigned to feature a different color than the one used on the *"Continuar"* and *"Explicação"* buttons. Given the fact that, since the beginning, the objective was to keep the current REAP.PT style and interface layout, alternatives to the placement of the buttons and other assets, or additional information to be displayed in the welcome page (apart from the instructions on how to "ideally" carry out the exercise solving task) were not considered.

Diacritical Marks

After submitting the puzzles, some users expected the system to perform in similar fashion to what occurs in traditional crosswords puzzles. In these, cells that contain a word with a diacritical mark, have said mark removed for the intersecting word. For example, given the word *"tédio"* (boredom), the cell with the <é> letter will act as a normal <e> for an intersecting word, such as *"nuvem"* (cloud). This is not the case in the current implementation, where letters with diacritical marks do not undergo any modifications for the intersecting words. For example, if the word *"adver**sá**rios"* (adversaries), is chosen, the syllable "-sá" will also be present in the corresponding intersecting word, such as, for example, *"colap**sá**vel"* (collapsible). That is not the case in traditional crossword puzzles, where words containing the syllable "-sa", such as *"anali**sá**vel"* (analysable) or *"pesqui**sa**das"* (surveyed), would be valid intersections. At one point, this feature was planned. However, given the number of available words and intersections, it was dropped because it did not have a major impact on the number of puzzles to be

generated.

Questionnaire redirection

One user noted that after the questionnaire was filled and submitted, he or she should have been redirected to the index page of the puzzle. This was not done because most users were not interested in solving additional exercises right after filling out the questionnaire. However, once the evaluation stage was finished, this issue was addressed by removing the questionnaire button from the feedback page.

5.3 Derivation Exercise Results

5.3.1 Crowd-sourced Test Results

Similar to what happened during the Crossword puzzle evaluation, a record was kept of every Derivation puzzle that was submitted for correction. In total, 570 puzzles were submitted.

		Exception			
		Easy	Medium	Hard	Total
Prefix	Easy	193	28	23	244
	Medium	51	50	13	114
	Hard	40	34	138	212
	Total	284	112	174	570

Table 5.6: Total number of submitted derivation puzzles by difficulty level.

The partial totals corresponding to each difficulty level can be found in Table 5.6. The details of the results per prefix (level, average exception level, average score, average hints and the total of submitted puzzles) can be found in Appendix C. As expected, the default combination, which corresponds to easy prefixes and easy exceptions, was the preferred one (33,9%). However, contrary to what happened on the Crossword puzzle exercise, apart from the hardest level possible (24,2%), no other particular combination was preferred. It is possible that most users chose to solve puzzles belonging to the easier difficulty levels and once they grew accustomed to the objective of the puzzle and its interface, they opted to increase the difficulty levels as much as possible.

The score ratios for each difficulty level can be found in Table 5.7. Once again, these values correspond to the score achieved by the users over the maximum possible score for each puzzle. The main

		Exceptions		
		Easy	Medium	Hard
Prefix	Easy	73.58%	83.93%	89.13%
	Medium	78.43%	68.00%	63.54%
	Hard	58.13%	58.82%	65.04%

Table 5.7: Average score of the submitted derivation puzzles, according to the difficulty levels.

difference between these values and the ones concerning the Crossword puzzle exercise, observable in Table 5.4, is the fact that there are no values under 50%. Besides this, the increase in difficulty levels does not necessarily translate into lower values. In fact, the opposite occurs when the difficulty level of the prefix is "easy" (first line), for example. However, if we consider the medium difficulty level prefixes (second line), a slight decrease can be found, which is apparently in contradiction with the easy level prefixes. In all, the increase in difficulty did not had such a strong impact in the users's scores in this game as it had in the Crossword puzzle.

		Exception			
		Easy	Medium	Hard	Total
Prefix	Easy	39	1	1	41
	Medium	4	11	2	17
	Hard	2	0	21	23
	Total	45	12	24	81

Table 5.8: Number of used hints for the submitted derivation puzzles, according to the difficulty levels.

Most users opted not to use any hints while solving the puzzles. In fact, the average number of used hints per submitted puzzle is just 0.14. In comparison to the Crossword puzzle exercise, it is possible that the users took more chances when submitting the puzzle, instead of making sure their answers were the correct ones. In the Crossword puzzle exercise, users had support features which allowed the puzzle to be "corrected" before submission (color), and automatically completed (hint). Because these puzzles took longer to complete, it is also possible that the users felt the need to use these features more often, in order to make sure the time spent during the process of solving the puzzles was not wasted. Given the fact that multiple-choice exercises are not particularly time-consuming, help features such as the one available in this exercise can be seen as a "secondary" feature, rather than an essential one.

5.3.2 Questionnaire Results

The questionnaire results for the derivation exercise, can be found in figures 5.12 to 5.18. In total, 53 users filled-out the form in question. Their average age was 25.5, ranging from 18 to 57 years old. 75% of the users had completed college education, with the remaining 25% having graduated high school.

Results for the third question, *"O português é a sua língua materna?"* (Is Portuguese your mother tongue?), were once again omitted because all inquired users were Portuguese native speakers.

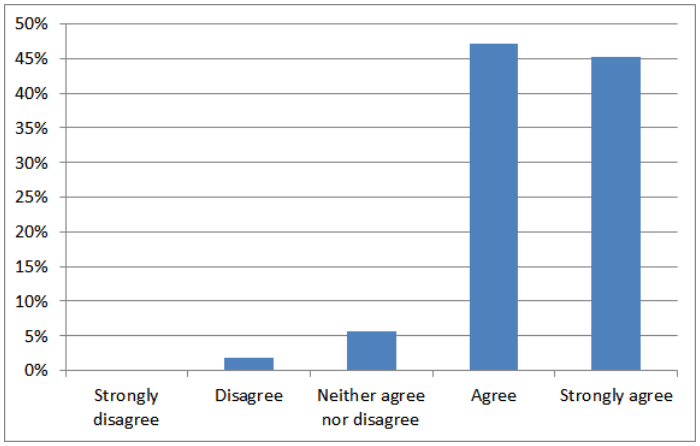

Figure 5.12: Derivation Questionnaire results for the statement "The system is easy to use.".

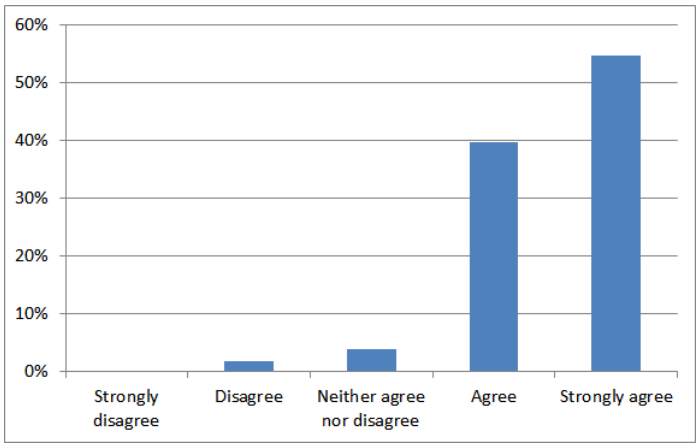

Figure 5.13: Derivation Questionnaire results for the statement "I understood the goal quickly.".

Results for the statement "The system is easy to use." can be found in Figure 5.12. 47% of the users agreed with this statement and 45% strongly agreed with it, totalling 92% positive answers. In fact, only one person disagreed with it. Similar percentages were also achieved in the question relative to the objective of the exercise ("I understood the goal quickly."), as can be seen in Figure 5.13. Once again, only one user disagreed with the statement. The majority of the users (55%) strongly agreed with it and 39% just agreed with it. In total, these correspond to 94% of the inquired population.

Concerning the matter of the exercises being too easy, the users did not reach a consensus, as the results in Figure 5.14 show. Even though only 25% agreed or strongly agreed with the statement in question, the majority of the users (39%) chose not to commit to an answer, that is, neither agreeing or

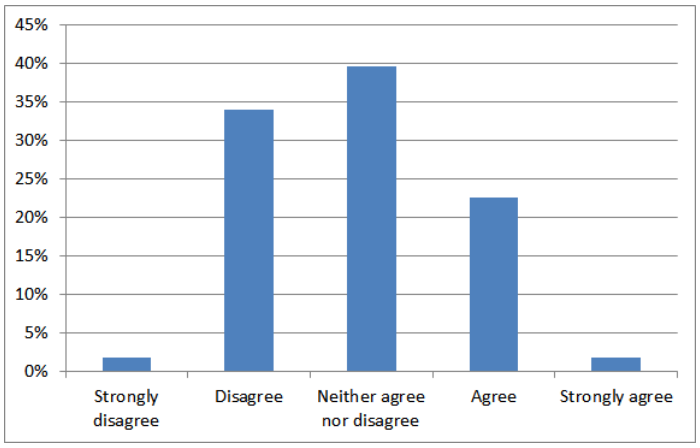

Figure 5.14: Derivation Questionnaire results for the statement "The exercises are too easy".

disagreeing with it. On the other hand, 36% of the users disagreed with the statement. This result can be related to the relatively high scores obtained in this game, as shown in Table 5.7. When compared with the Crossword puzzle, this perception from the users hints at the more complex metalinguistic awareness and grammatical processes involved in this game's solving.

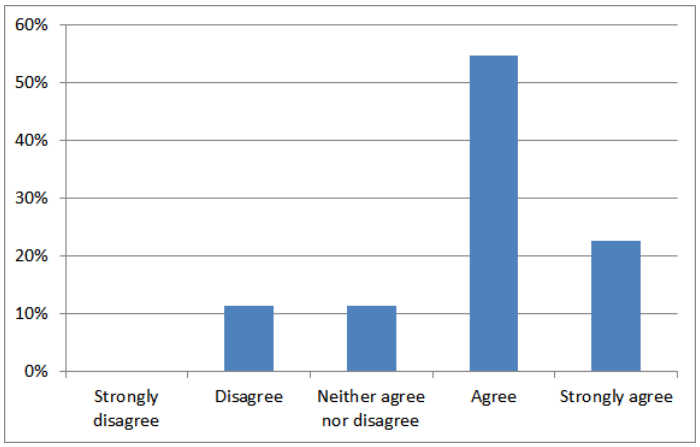

Figure 5.15: Derivation Questionnaire results for the statement "I noticed there were differences across the difficulty levels.".

Interestingly, however, 77% of the users at least agreed with the statement "I noticed there were differences across the difficulty levels.", as it can be seen in Figure 5.15. These results, combined with the ones in Figure 5.14, seem to suggest that even tough the users noticed the differences between difficulty levels, they were not impossible or extremely hard to solve, even at the highest possible combination of difficulty levels.

Regarding the usefulness of the exercise, more precisely the statement "I learnt something using the

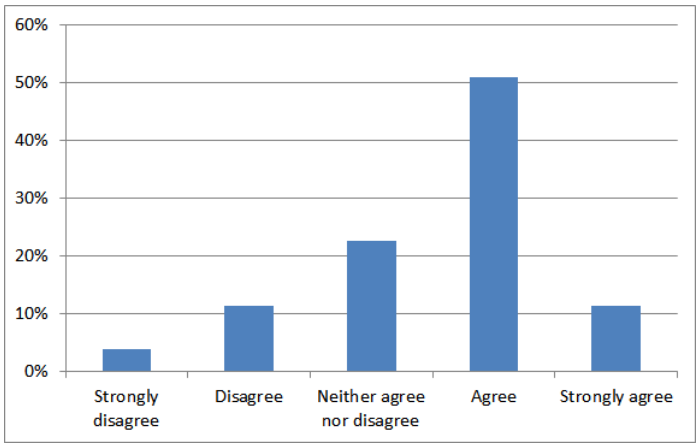

Figure 5.16: Derivation Questionnaire results for the statement "I learnt something using the system.".

system.", results shown in Figure 5.16, indicate that 51% of the users agreed with the statement and 11% strongly agreed with it, totalling 62% of the answers, while only 15% stated the opposite; 22% neither agreed nor disagreed. Compared to the Crossword puzzle, these results show that users claimed to have learnt more by using this game.

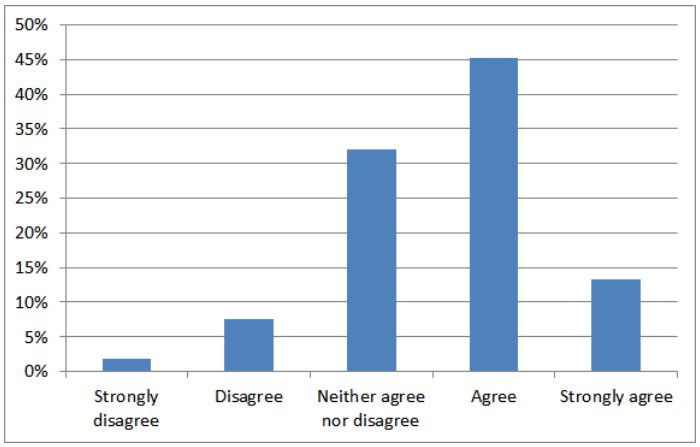

Figure 5.17: Derivation Questionnaire results for the statement "I had fun using the system.".

Results for the statement "I had fun using the system." can be found in Figure 5.17. This question concerns the Satisfaction factor and its results were once again positive. The percentage of the users who, at least, agreed with this statement was exactly the same for both exercises (58%). For this specific exercise, only 9% disagreed with the sentence.

Finally, 79% of the users claimed to be at least "somewhat satisfied" with the exercise, as can be seen in Figure 5.18. In fact, no users claimed to be "very dissatisfied" and only two users stated that they were "somewhat dissatisfied" with the system, which, once again, validates the approach followed

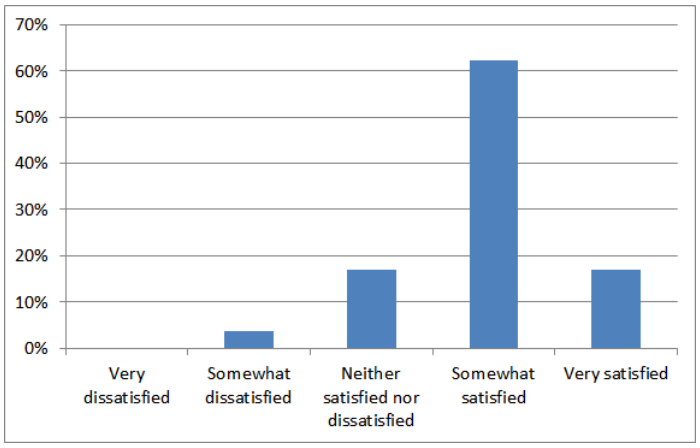

Figure 5.18: Derivation Questionnaire results for the statement "Global system appreciation.".

in this project.

5.3.3 Questionnaire Comments

The free-form text box in the final question of both questionnaires, allowed for personalized feedback to be posted. These suggestions include possible modifications to the exercise and other aspects the users thought should be improved or altered in some way.

Lack of examples

Several users complained about the lack of sentences, in the feedback screen showing the words from the exercise in use. The majority considered that these sentences would have been especially useful whenever they included words they did not know about. Since only one resource, the CETEMPúblico corpus, is used for this task, it is normal for some words not to have at least one sentence containing them. While a 100% coverage rate regarding the entirety of the exceptions and the distractors is difficult to achieve, given the total number of words, the inclusion of additional corpora should help increase this rate.

Adaptive difficulty levels

One of the most frequent suggestions was the inclusion of a feature that automatically increased or decreased the difficulty levels of the exercises, depending on the performance of the users while solving the puzzles. Instead of having the user manually altering the difficulty levels after, successfully or unsuccessfully, solving a sequence of puzzles, the system would suggest these alterations, allowing for

a more interactive and streamlined experience. This type of feature was actually considered during the development stage of this game, as it corresponds to the general framework of REAP.PT, adapting to the user's performance. However, it was eventually dropped in favour of maintaining the "standard" game-play style of the already available syntactical exercises, more specifically, the multiple-choice ones, which are stable (a difficulty level for each gaming session).

Dictionary integration

Similar to what happened during the solving process of the Crossword puzzle exercise, some users expected a larger variety of help features. A recurrently suggested one was the inclusion, as clues, of the definitions of the words, extracted from on-line dictionaries such as *Infopédia* or Priberam. While the definitions could help the users understand the meaning behind the words, they typically contain information on how the word is formed which, in the case of this exercise, defeats the purpose of trying to identify which words are synchronically formed by prefix addition and which are not. In other words, though a morphological process may underlie the exceptions, and the dictionary would indicate it, the purpose of the exercise was to recognize that the resulting form can no longer be associated to a base form, something that only the comparison of the definitions would clarify. Besides, this unwanted information on base prefixed form meaning is not limited to a specific place, as it was the case of the definitions in the Crossword puzzle exercise, manual correction of the extracted definitions would be required. Given the number of exceptions and distractors in this exercise, this process would take too much time and, for that reason, this help feature was not implemented. The inclusion of the definitions of the prefixes in the help screen, and the fact that users are not penalized score-wise when they access these definitions, offer an alternative to the definition of the words in their entirety.

Prefix application

One user, familiar with the already available syntactic exercises of the REAP.PT system, suggested the implementation of a new game, using the resources of this Derivation puzzle exercise. Similar to what happens with the Lexical Mahjong exercise, users would be asked to match blocks consisting of valid combinations of prefixes and words. This type of puzzle is actually a variation of the third exercise found in Section 3.3.2 of this document. Regarding the implementation, this would be done by either dragging and dropping prefixes onto the words, or by using the same interface used by the Lexical Mahjong exercise (clicking on the corresponding tiles). During the planning stage, an exercise similar to the one proposed, using both prefixes and suffixes was considered, but eventually it was dropped in favour of the followed multiple-choice approach.

6 Conclusion and Future Work

6.1 Final Remarks

Given the current expansion and ongoing integration of ICALL systems in many languages, it is important that new and innovative approaches are taken when building new tools to be integrated in these systems. Not only do these systems offer a larger variety of choice when studying various resources and aspects of languages, but they also may represent a valid alternative to traditional exercises, while affecting the way teachers interact with their students, increasing motivation and improving their language learning curves.

The work developed in this dissertation consists of various valuable assets, which can be applied and easily refitted to serve several of these purposes. The fact that both exercises offer several combinations of difficulty levels means that they can be used by a wide range of people, regardless of their age and language proficiency level. It also helps keeping higher motivation on students, as the difficulty level can be adjusted to their progress.

Even though some unexpected difficulties were unearthed, during the development stage, which mainly concerned the YAH hyphenator, the positive results obtained from hyphenating both the P-AWL and the alternative corpus prove that this tool can still be used effectively and with a reasonable degree of confidence. The correction of the errors that existed in the original version of this tool, allows for future updates and revisions to the current implementation, to focus solely on the expansion and revision of the applicable rules, which would the achievement of even higher success rates, regardless of the processed corpora .

The Crossword puzzle exercise offers a unique and original twist on the traditional crossword puzzles, opting to focus on word formation by requiring users to correctly divide the words in syllables. Besides this, identical to what happens in traditional crossword puzzles, the vocabulary of the users is also tested and expanded, as the clues given to uncover the hidden words are extracted from real dictionaries. Furthermore, due to its modular architecture, which has the linguistic resources (the definitions and the list of divided words) completely separated from the source code, it can be easily adapted to foreign languages. This allows for alternative assets, namely lists of different definitions and words, to be directly incorporated in the exercise framework.

Similarly, the Derivation exercise can also be easily updated and adapted to fit different requirements. Once more, the linguistic resources are also separated from the code, which allows for a completely new different types of multiple-choice exercises to be implemented. The parameters used for choosing the distractors and for sorting assets among different difficulty levels, can also be easily changed.

Finally, the fact that both systems present the user with personalized feedback depending on what they answered allows for both correct and incorrect answers to be compared and observed simultaneously. This is especially important in the Derivation exercise, where the real sentence examples offer precious information on how and when to apply the words in question. It also means that instead of simply informing the users of their performance's result, an extra step is taken in order to give them more information on their errors, and to prevent them from repeating the same errors in the future.

6.2 Future Work

In this section, corrections are proposed for some of the problems found during the evaluation stage.

YAH extension

A redesign of the YAH's architecture would not only allow it to be ported to other languages, but it would increase its effectiveness as well. In its current implementation, the rules are incorporated in the source files, which makes the process of editing them, not trivial, since it would involve combining linguistic expertise with some programming skills. By separating the rules from the source code, and by making their syntax more user-friendly, other stakeholders could take part in the process of improving the overall quality of the hyphenator. Besides, this would allow for other languages to be supported alongside Portuguese, as the system would not be limited to a specific set of rules. Finally, alongside the already available word beginning rules and other "generic" rules, the inclusion of "word-ending" rules and "hybrid" rules, capable of handling both mask and word characters, would increase the versatility and flexibility of this application, as well as reduce the number of exceptions and unique cases which have to be dealt with accordingly.

Resource maintenance

The resources from the derivation exercise would benefit greatly from having human maintenance in the long term. Since the process of generating puzzles is not particularly complex, some exercises may end up being too easy because the appropriate distractors are not always chosen. For example, using the prefix "bi-", one of the generated exercises consisted of the target word *"bilreira"* (bobbin lace artisan),

and the distractors *"bicampeonato"* (bi-championship), *"bipólo"* (dipole) and *"bicelular"* (bicellular). Given its morphology, *bilreira* should not be compared with words with the prefix "bi-" (as no base word in Portuguese starts with "lr-".

Some restrictions were implemented in order to prevent prefixes from using the resources belonging to other prefixes. For example, even tough both "in-" and "inter-" are used to generate exercises, the prefix "in-" is not allowed to use the "inter-" distractors, even tough they all begin with the substring "in". These were not extensively listed and were only added to prevent conflicting sets of resources from being used. With this in mind, the process of (once again) manually correcting the lists of words (exceptions and distractors) and the addition of said rules, would improve the overall quality of the generated exercises.

Alternative sources

Both exercises would benefit from having their resources updated in the long term. On one hand, concerning the crossword puzzle exercise, alternative sources, such as the *Priberam*[1] on-line dictionary, could also be used to generate new, different clues for the remaining words (apart from the adverbs alone, as it is now used). It would be interesting to let the users choose their favourite lexical resource. The inclusion of new dictionary definitions would also allow for a larger number of clues for each word.

On the other hand, other corpora and tools should be incorporated in the derivation exercise. While the Ensinador[2] tool from Linguateca performs well under normal circumstances, during the evaluation stage, its server proved not to be able to handle a substantial amount of requests. Besides this, even though CETEMPúblico (Santos & Rocha, 2001) is a sizeable corpus, it does not contain examples for every word used in this exercise. The use of a web-crawler, such as the one in the Daily REAP.PT component, is a possible alternative for dealing with this issue. An alternative to the extraction of real sentences would be the presentation of definitions for both the the lemma and the pseudo-prefixed word.

6.3 Contributions

The main contributions of this thesis are:

- A new updated version of the YAH hyphenator, capable of correctly dividing a bigger number of words;

[1] `http://www.priberam.pt/dlpo/` (last visited in August 2013)
[2] `http://www.linguateca.pt/Ensinador/index.pl?corpus=CETEMPUBLICO` (last visited in August 2013)

- A new Crossword Puzzle game, focusing on syllables instead of individual letters. The fact that all linguistic resources are independent from the source code, allows for this exercise to be easily adapted to other languages;
- A new syntactic game focusing on prefixal derivation, featuring dynamic feedback. Once again, its modular architecture modifier allows for other linguistic resources to be applied and used;
- Various graphical resources (buttons and other elements) created for these exercises but reusable in future additions to the REAP.PT system.

Bibliography

Amaral, L. A. & D. Meurers (2011, January). On using intelligent computer-assisted language learning in real-life foreign language teaching and learning. *ReCALL 23*(1), 4–24.

Baptista, J., N. Costa, J. Guerra, M. Zampieri, M. Cabral, & N. Mamede. P-AWL: Academic Word List for Portuguese. In T. Pardo, A. Branco, A. Klautau, R. Vieira, & V. Lima (Eds.), *Computational Processing of the Portuguese Language*, Volume 6001 of *Lecture Notes in Computer Science*, pp. 120 – 123. Springer Berlin / Heidelberg.

Collins-Thompson, K. & J. P. Callan (2005). Predicting reading difficulty with statistical language models. *Journal of the American Society for Information Science and Technology 56*(13), 1448–1462.

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

Filipe, H. (2013). Redesign da Interface de Utilizador do REAP.PT. Master's Thesis, Universidade do Algarve.

Freitas, T. (2012). Syntactic REAP.PT - Exercises on Clitic Pronouning. Master's Thesis, Instituto Superior Técnico - Universidade Técnica de Lisboa.

Levenshtein, V. (1966). Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady 10*, 707.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology 22*(140), 1–55.

Lopes, M. d. C. V. (2011). *Gramática da Língua Portuguesa 2.° Ciclo do Ensino Básico 5.°/6.° Anos*. Plátano Editora.

Lopes, M. d. C. V. (2012). *Domínios-Gramática da Língua Portuguesa, Exercícios de Aplicação*. Plátano Editora.

Lund, A. M. (2001, October). Measuring Usability with the USE Questionnaire. *STC Usability SIG Newsletter 8*(2).

Mamede, N., J. Baptista, C. Diniz, & V. Cabarrão (2012). STRING: An Hybrid Statistical and Rule-Based Natural Language Processing Chain for Portuguese. In *Proceedings of the 12th international conference on Computational Processing of the Portuguese Language - Demo sessions http://www.propor2012.org/demos/DemoSTRING.pdf*, PROPOR 2012.

Marques, C. (2011). Syntactic REAP.PT. Master's Thesis, Instituto Superior Técnico - Universidade Técnica de Lisboa.

Marrafa, P., R. Amaro, S. Mendes, S. Lourosa, & R. Chaves (2006). TemaNet - WordNets Temáticas do Português: Alimentação, Arte, Comunicação, Desporto, Educação, Geografia, Habitação, Relações Humanas, Saúde, Seres Vivos, Transportes e Turismo. `http://www.instituto-camoes.pt/temanet`.

Marujo, L. (2009). REAP em Português. Master's Thesis, Instituto Superior Técnico - Universidade Técnica de Lisboa.

Meurers, D., R. Ziai, L. Amaral, A. Boyd, A. Dimitrov, V. Metcalf, & N. Ott (2010). Enhancing authentic web pages for language learners. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, IUNLPBEA '10, Stroudsburg, PA, USA, pp. 10–18. Association for Computational Linguistics.

Moreira, J. E., M. M. Michael, D. Da Silva, D. Shiloach, P. Dube, & L. Zhang (2007). Scalability of the Nutch Search Engine. In *Proceedings of the 21st annual International Conference on Supercomputing*, ICS '07, New York, NY, USA, pp. 3–12. ACM.

Oliveira, H., D. Santos, P. Gomes, & N. Seco (2008). Papel: A dictionary-based lexical ontology for portuguese. In A. Teixeira, V. Lima, L. Oliveira, & P. Quaresma (Eds.), *Computational Processing of the Portuguese Language*, Volume 5190 of *Lecture Notes in Computer Science*, pp. 31–40. Springer Berlin Heidelberg.

Paulo, S., L. C. Oliveira, C. Mendes, L. Figueira, R. Cassaca, C. Viana, & H. Moniz (2008). DIXI — A Generic Text-to-Speech System for European Portuguese. In *Proceedings of the 8th international conference on Computational Processing of the Portuguese Language*, PROPOR 2008, Berlin, Heidelberg, pp. 91–100. Springer-Verlag.

Santos, D. & P. Rocha (2001). Evaluating CETEMPúblico, a free resource for Portuguese. In *Proceedings of the 39th Annual Meeting on the Association for Computational Linguistics*, ACL '01, Morristown, NJ, USA, pp. 450–457. Association for Computational Linguistics.

Silva, A. (2011). Pictorial REAP.PT. Master's Thesis, Instituto Superior Técnico - Universidade Técnica de Lisboa.

Sweetser, P. & P. Wyeth (2005). Gameflow: a model for evaluating player enjoyment in games. *Computers in Entertainment 3*(3), 3.

Trancoso, I., T. Pellegrini, A. F. Silva, R. Correia, N. J. Mamede, & J. Baptista (2012, November). Learning portuguese with speech technologies. In *IBERSPEECH*, pp. 411–414.

Vicente, A. (2013). LexMan: um Segmentador e Analisador Morfológico com Transdutores. Master's Thesis, Instituto Superior Técnico - Universidade Técnica de Lisboa.

I

Appendices

A Modified YAH Rules

```
// Rules aplicable to the beginning of the word.

word_beginning_rules.push_back(Rule("bi", 2));
word_beginning_rules.push_back(Rule("cláu", 4));
word_beginning_rules.push_back(Rule("clás", 3));
word_beginning_rules.push_back(Rule("cri", 3));
word_beginning_rules.push_back(Rule("cue", 2));
word_beginning_rules.push_back(Rule("coin", 2));
word_beginning_rules.push_back(Rule("prio", 3));
word_beginning_rules.push_back(Rule("proi", 3));
word_beginning_rules.push_back(Rule("proí", 3));
word_beginning_rules.push_back(Rule("rees", 2));
word_beginning_rules.push_back(Rule("rein", 2));
word_beginning_rules.push_back(Rule("reo", 2));
word_beginning_rules.push_back(Rule("sta", 3));
word_beginning_rules.push_back(Rule("subl", 3));
word_beginning_rules.push_back(Rule("vié", 2));
word_beginning_rules.push_back(Rule(std::string(1, type_1_consonant_mask)
    += type_2_consonant_mask, -1, true)); //ex:pl
word_beginning_rules.push_back(Rule(std::string(1, type_1_consonant_mask)
    += consonant_mask, -1, true)); //ex:pc

// Rules aplicable to the rest of the word.

rules.push_back(Rule("aí", 1));
rules.push_back(Rule("air", 1));
rules.push_back(Rule("ch", 0));
rules.push_back(Rule("cção", 1));
rules.push_back(Rule("ção", 0));
rules.push_back(Rule("ções", 0));
rules.push_back(Rule("een", 1));
rules.push_back(Rule("eí", 1));
rules.push_back(Rule("gua", 0));
rules.push_back(Rule("gui", 0));
rules.push_back(Rule("gue", 0));
rules.push_back(Rule("guo", 0));
rules.push_back(Rule("io", 1));
rules.push_back(Rule("lh", 0));
rules.push_back(Rule("nh", 0));
rules.push_back(Rule("oe", 1));
rules.push_back(Rule("oo", 1));
rules.push_back(Rule("pção", 1));
rules.push_back(Rule("pções", 1));
rules.push_back(Rule("pt", 0));
rules.push_back(Rule("que", 0));
rules.push_back(Rule("rr", 0));
```

```
rules.push_back(Rule("ssões", 0));
rules.push_back(Rule("sões", 0));
rules.push_back(Rule("ss", 0));
rules.push_back(Rule("uí", 1));

rules.push_back(Rule(std::string(1, type_1_consonant_mask)
    +=type_2_consonant_mask, 0, true)); //ex:pl
rules.push_back(Rule(std::string(1, type_i_vowel_mask)
    +=vowel_mask, 1, true)); //ex: iv
rules.push_back(Rule(std::string(1, type_u_vowel_mask)
    +=vowel_mask, 1, true)); //ex: uv
rules.push_back(Rule(((std::string(1, consonant_mask)
    +=type_u_vowel_mask)+=type_i_vowel_mask)+=consonant_mask, 0, true)); //ex:cuic, divided again at 2
rules.push_back(Rule((std::string(1, consonant_mask)
    +=type_u_vowel_mask)+=vowel_mask, 0, true)); //ex:cuv, divided again at 2
rules.push_back(Rule(std::string(1, consonant_mask)
    +=type_u_vowel_mask, 0, true)); //ex: cu
rules.push_back(Rule((std::string(1, consonant_mask)
    +=type_i_vowel_mask)+=vowel_mask, 0, true)); //ex:civ, divided again at 2
rules.push_back(Rule((std::string(1, consonant_mask)
    +=type_i_vowel_mask)+=type_u_vowel_mask, 0, true)); //ex:ciu, divided again at 2
rules.push_back(Rule(std::string(1, consonant_mask)
    +=type_i_vowel_mask, 0, true)); //ex:ci
rules.push_back(Rule(std::string(1, consonant_mask)
    +=vowel_mask, 0, true)); //ex: cv
rules.push_back(Rule(((std::string(1, type_2_consonant_mask)
    +=type_u_vowel_mask)+=type_i_vowel_mask)
    +=consonant_mask, 0, true)); //ex:luic, divided again at 2
rules.push_back(Rule(((std::string(1, type_2_consonant_mask)
    +=type_u_vowel_mask)+=type_i_vowel_mask)
    +=type_2_consonant_mask, 0, true)); //ex:luil, divided again at 2
rules.push_back(Rule(((std::string(1, type_2_consonant_mask)
    +=vowel_mask)+=type_1_consonant_mask)
    +=consonant_mask, 0, true)); //ex:lvpc, divided again at 3
rules.push_back(Rule((std::string(1, type_2_consonant_mask)
    +=type_i_vowel_mask)+=vowel_mask, 0, true)); //ex:liv, divided again at 2
rules.push_back(Rule((std::string(1, type_2_consonant_mask)
    +=type_i_vowel_mask)+=type_u_vowel_mask, 0, true)); //ex:liu, divided again at 2
rules.push_back(Rule(std::string(1, type_2_consonant_mask)
    +=type_u_vowel_mask, 0, true)); //ex:lu
rules.push_back(Rule(std::string(1, type_2_consonant_mask)
    +=type_i_vowel_mask, 0, true)); //ex:li
rules.push_back(Rule(std::string(1, type_2_consonant_mask)
    +=vowel_mask, 0, true)); //ex:lv
rules.push_back(Rule(((std::string(1, type_1_consonant_mask)
    +=type_i_vowel_mask)+=type_1_consonant_mask)
```

```
    +=type_1_consonant_mask, 0, true)); //ex:pipp, divided again at 3
rules.push_back(Rule(((std::string(1, type_1_consonant_mask)
    +=type_u_vowel_mask)+=type_i_vowel_mask)
    +=type_u_vowel_mask, 0, true)); //ex:puiu, divided again at 3
rules.push_back(Rule(((std::string(1, type_1_consonant_mask)
    +=type_u_vowel_mask)+=type_i_vowel_mask)
    +=consonant_mask, 0, true)); //ex:puic, divided again at 2
rules.push_back(Rule(((std::string(1, type_1_consonant_mask)
    +=type_u_vowel_mask)+=type_i_vowel_mask)
    +=type_2_consonant_mask, 0, true)); //ex:puil, divided again at 2
rules.push_back(Rule((std::string(1, type_1_consonant_mask)
    +=type_i_vowel_mask)+=vowel_mask, 0, true)); //ex:piv, divided again at 2
rules.push_back(Rule((std::string(1, type_1_consonant_mask)
    =type_i_vowel_mask)+=type_u_vowel_mask, 0, true)); //ex:piu, divided again at 2
rules.push_back(Rule((std::string(1, type_1_consonant_mask)
    +=type_u_vowel_mask)+=vowel_mask, 0, true)); //ex:puv, divided again at 2
rules.push_back(Rule(std::string(1, type_1_consonant_mask)
    +=type_u_vowel_mask, 0, true)); //ex:pu
rules.push_back(Rule(std::string(1, type_1_consonant_mask)
    +=type_i_vowel_mask, 0, true)); //ex:pi
rules.push_back(Rule(std::string(1, type_1_consonant_mask)
    +=vowel_mask, 0, true)); //ex:pv
rules.push_back(Rule((std::string(1, type_u_vowel_mask)
    +=type_i_vowel_mask)+=consonant_mask, 1, true)); //ex:uic
rules.push_back(Rule((std::string(1, type_u_vowel_mask)
    +=type_i_vowel_mask)+=type_u_vowel_mask, 1, true)); //ex:uiu, divided again at 2

//Exceptions and Word Endings

pos = result.find("ie"); if (pos != string::npos) {result.replace(pos, 2, "i=e");}//exception
pos = result.find("iê"); if (pos != string::npos) {result.replace(pos, 3, "i=ê");}//exception
pos = result.find("=í"); if (pos != string::npos)
                                                {
                                                        size_t pos1 = result.find("=ís");
                                                        if (pos1 != string::npos){}
                                                        else{result.replace(pos, 3, "=í=");}//exception
                                                }
pos = result.find("oé"); if (pos != string::npos) {result.replace(pos, 3, "o=é");}//exception
pos = result.find("õ=e"); if (pos != string::npos) {result.replace(pos, 4, "õe");}//exception
pos = result.find("e=u"); if (pos != string::npos) {result.replace(pos, 3, "eu");}//ending
pos = result.find("uir"); if (pos != string::npos) {result.replace(pos, 3, "u=ir");}//ending
pos = result.find("gu=ir"); if (pos != string::npos) {result.replace(pos, 5, "guir");}//ending override
pos = result.find("uí"); if (pos != string::npos) {result.replace(pos, 3, "u=í");}//exception
pos = result.find("uo"); if (pos != string::npos) {result.replace(pos, 2, "u=o");}//exception
pos = result.find("tui="); if (pos != string::npos) {result.replace(pos, 4, "tu=i=");}//exception
pos = result.find("bui="); if (pos != string::npos) {result.replace(pos, 4, "bu=i=");}//exception
```

```
pos = result.find("qu="); if (pos != string::npos) {result.replace(pos, 3, "qu");}//exception
```

B
Prefix Data

Prefix	#Distractors	#Exceptions	Level
aero	34	23	3-Hard
ante	42	34	3-Hard
anti	973	36	2-Medium
arqui	28	18	3-Hard
auto	681	60	2-Medium
bem	13	13	3-Hard
bi	115	208	1-Easy
contra	189	100	2-Medium
des	2878	962	1-Easy
dis	15	200	3-Hard
eco	45	24	3-Hard
eletro	63	19	3-Hard
endo	10	25	3-Hard
ex	31	371	3-Hard
extra	21	44	3-Hard
hidro	38	28	3-Hard
hiper	287	19	3-Hard
hipo	15	32	3-Hard
in	1123	992	1-Easy
inter	395	154	1-Easy
mal	78	111	2-Medium
meta	61	43	2-Medium
micro	262	13	3-Hard
mini	284	27	2-Medium
mono	123	63	2-Medium
multi	261	34	2-Medium
neo	199	14	3-Hard
paleo	24	14	3-Hard
para	77	98	2-Medium
peri	10	41	3-Hard
piro	12	12	3-Hard
poli	71	68	2-Medium
psico	42	18	3-Hard
quadri	27	23	3-Hard
radio	46	41	3-Hard
re	1696	1667	1-Easy
retro	29	23	3-Hard
semi	464	28	2-Medium
sobre	267	98	2-Medium
socio	42	62	3-Hard
sub	558	173	1-Easy
super	674	41	2-Medium
tele	241	49	2-Medium
tetra	12	11	3-Hard
trans	127	214	1-Easy
tri	54	195	2-Medium
ultra	218	18	3-Hard
uni	42	62	3-Hard

C Derivation Exercise Results

Prefix	Level	Avg. Exception Level	Avg. Score	Avg. Hints	Submitted Puzzles
aero	3	2,50	34,85%	0,00	12
ante	3	2,63	75,56%	0,00	8
anti	2	1,57	84,00%	0,00	7
arqui	3	2,33	81,25%	0,11	9
auto	2	1,75	73,33%	0,38	8
bem	3	2,23	55,88%	0,00	13
bi	1	1,30	67,92%	0,00	23
contra	2	1,20	81,25%	0,00	10
des	1	1,28	66,29%	0,21	39
dis	3	2,56	68,00%	0,00	9
eco	3	2,71	55,00%	0,00	7
electro	3	3,00	67,86%	0,43	7
endo	3	2,43	73,68%	0,00	7
ex	3	3,00	91,67%	0,17	6
extra	3	2,33	89,58%	0,00	9
hidro	3	2,47	67,07%	0,40	15
hiper	3	2,67	56,86%	0,00	9
hipo	3	1,00	31,25%	0,50	4
in	1	1,26	80,23%	0,16	38
inter	1	1,31	76,29%	0,31	42
mal	2	1,80	84,21%	0,40	5
meta	2	1,50	85,71%	0,38	8
micro	3	2,43	47,37%	0,00	7
mini	2	2,00	33,33%	0,00	6
mono	2	1,43	50,00%	0,00	7
multi	2	1,70	54,86%	0,20	10
neo	3	3,00	66,67%	0,00	3
paleo	3	2,00	74,29%	0,00	7
para	2	2,31	75,00%	0,00	13
peri	3	2,57	100,00%	0,00	7
piro	3	2,00	64,44%	0,11	9
poli	2	1,50	82,65%	0,07	14
psico	3	2,63	48,89%	0,00	8
quadri	3	2,63	37,78%	0,00	8
radio	3	2,36	69,49%	0,00	11
re	1	1,44	94,32%	0,03	36
retro	3	2,67	43,14%	0,00	9
semi	2	1,38	77,78%	0,00	8
sobre	2	2,00	70,83%	0,00	6
socio	3	1,75	73,68%	0,00	4
sub	1	1,21	78,77%	0,27	33
super	2	2,00	100,00%	0,00	2
tele	3	1,50	100,00%	0,00	2
tetra	3	2,36	55,93%	0,55	11
trans	1	1,30	76,32%	0,12	33
tri	2	1,50	44,64%	0,75	8
ultra	3	2,71	45,00%	0,43	7
uni	3	2,83	82,86%	0,00	6

D Crossword Questionnaire

Palavras Cruzadas

Obrigado pela sua participação neste projeto. A sua opinião é muito importante para nós.

Por favor preencha o seguinte questionário:

*Obrigatório

1. **1. Qual a sua idade?** *
   (Insira um número)

   .....

2. **2. O português é a sua língua materna?** *
   Marcar apenas uma oval.

   - Sim
   - Não

3. **3. Qual o seu grau de escolaridade?** *
   Marcar apenas uma oval.

   - 1º ciclo do ensino básico
   - 2º ciclo do ensino básico
   - 3º ciclo do ensino básico
   - Ensino secundário
   - Ensino superior

4. **4. O sistema é fácil de utilizar.** *
   Marcar apenas uma oval.

   - Discordo totalmente
   - Discordo
   - Não concordo, nem discordo
   - Concordo
   - Concordo totalmente

5. **5. Percebi rapidamente o objetivo.** *
   Marcar apenas uma oval.

   - Discordo totalmente
   - Discordo
   - Não concordo, nem discordo
   - Concordo
   - Concordo totalmente

6. **6. Depois de identificar a palavra, tive dificuldades em dividi-la.** *
Marcar apenas uma oval.

- Discordo totalmente
- Discordo
- Não concordo, nem discordo
- Concordo
- Concordo totalmente

7. **7. Os exercícios são demasiado fáceis.** *
Marcar apenas uma oval.

- Discordo totalmente
- Discordo
- Não concordo, nem discordo
- Concordo
- Concordo totalmente

8. **8. O feedback (retorno) dado pelo sistema é suficiente.** *
Marcar apenas uma oval.

- Discordo totalmente
- Discordo
- Não concordo, nem discordo
- Concordo
- Concordo totalmente

9. **9. Notei diferenças entre os graus de dificuldade** *
Marcar apenas uma oval.

- Discordo totalmente
- Discordo
- Não concordo, nem discordo
- Concordo
- Concordo totalmente

10. **10. Aprendi alguma coisa ao utilizar o sistema.** *
Marcar apenas uma oval.

- Discordo totalmente
- Discordo
- Não concordo, nem discordo
- Concordo
- Concordo totalmente

11. **11. Diverti-me ao utilizar o sistema.** *
Marcar apenas uma oval.

- Discordo totalmente
- Discordo
- Não concordo, nem discordo
- Concordo
- Concordo totalmente

12. **12. Qual a ajuda mais útil?** *
Marcar apenas uma oval.

- Colorir (O sistema de cores utilizado para distinguir as células corretas das erradas)
- Dica (O preenchimento automático de células específicas)
- Mais Definições (A apresentação de definições adicionais para as palavras em questão)

13. **13. Qual a ajuda menos útil?** *
Marcar apenas uma oval.

- Colorir (O sistema de cores utilizado para distinguir as células corretas das erradas)
- Dica (O preenchimento automático de células específicas)
- Mais Definições (A apresentação de definições adicionais para as palavras em questão)

14. **14. Apreciação global do sistema.** *
Marcar apenas uma oval.

- Muito insatisfeito
- Insatisfeito
- Nem satisfeito, nem insatisfeito
- Satisfeito
- Muito satisfeito

15. **Comentários Adicionais:**
De que gostou mais? E menos? O que mudaria? Como podemos melhorar o sistema?

.....

.....

.....

.....

.....

E

Derivation Questionnaire

Prefixação

Obrigado pela sua participação neste projeto. A sua opinião é muito importante para nós.

Por favor preencha o seguinte questionário:

*Obrigatório

1. **1. Qual a sua idade?** *
   (Insira um número)

   .....

2. **2. O português é a sua língua materna?** *
   Marcar apenas uma oval.

   - Sim
   - Não

3. **3. Qual o seu grau de escolaridade?** *
   Marcar apenas uma oval.

   - 1º ciclo do ensino básico
   - 2º ciclo do ensino básico
   - 3º ciclo do ensino básico
   - Ensino secundário
   - Ensino superior

4. **4. O sistema é fácil de utilizar.** *
   Marcar apenas uma oval.

   - Discordo totalmente
   - Discordo
   - Não concordo, nem discordo
   - Concordo
   - Concordo totalmente

5. **5. Percebi rapidamente o objetivo.** *
   Marcar apenas uma oval.

   - Discordo totalmente
   - Discordo
   - Não concordo, nem discordo
   - Concordo
   - Concordo totalmente

6. **6. Os exercícios são demasiado fáceis.** *
Marcar apenas uma oval.

- Discordo totalmente
- Discordo
- Não concordo, nem discordo
- Concordo
- Concordo totalmente

7. **7. Notei diferenças entre os graus de dificuldade** *
Marcar apenas uma oval.

- Discordo totalmente
- Discordo
- Não concordo, nem discordo
- Concordo
- Concordo totalmente

8. **8. Aprendi alguma coisa ao utilizar o sistema.** *
Marcar apenas uma oval.

- Discordo totalmente
- Discordo
- Não concordo, nem discordo
- Concordo
- Concordo totalmente

9. **9. Diverti-me ao utilizar o sistema.** *
Marcar apenas uma oval.

- Discordo totalmente
- Discordo
- Não concordo, nem discordo
- Concordo
- Concordo totalmente

10. **10. Apreciação global do sistema.** *
Marcar apenas uma oval.

- Muito insatisfeito
- Insatisfeito
- Nem satisfeito, nem insatisfeito
- Satisfeito
- Muito satisfeito

11. **Comentários Adicionais:**
De que gostou mais? E menos? O que mudaria? Como podemos melhorar o sistema?

.....

.....

.....

.....

.....

Com tecnologia
Google Drive