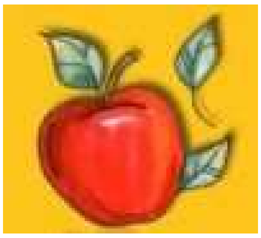
INSTITUTO SUPERIOR TÉCNICO
Universidade Técnica de Lisboa



# Automatic Question Generation for REAP.PT Tutoring System

**Rui Pedro dos Santos Correia**

Dissertation for obtaining the Master's Degree in
**Information Systems and Computer Engineering**

| | |
|---|---|
| Advisor: | Doutor Nuno João Neves Mamede |
| Co-advisor: | Doutora Isabel Maria Martins Trancoso |

**July 2010**

To my parents, Mário and Isabel.

To my sister, Joana,

# Acknowledgements

I would like to thank my advisors, Professor Nuno Mamede and Professor Isabel Trancoso, for all their guidance, discussion and motivation. Thank you also to all members of the REAP project at Carnegie Mellon University, namely Professor Maxine Eskenazi, Adam Skory, Gabriel Parent and Luís Marujo for their hospitality during my visit to Carnegie Mellon University.

I also want to thank all the members of L$^2$F, either for providing me essential tools for the development of this work or for helping me with discussion. Many thanks to my laboratory partners Miguel Bugalho, Wang Lin and Paula Vaz. Many thanks to Tiago Luís for all his help and endless availability.

I want to express my gratitude to the members of University of Algarve for their constant help, discussion and materials provided, essential to the execution of my work.

To FCT, for sponsoring this research under grant (INESC–ID multiannual funding) through the PIDDAC Program funds and the REAP.PT project.

Many thanks to all my family for the wise advices, and for always believing in me.

Finally, to my friends, Teresa Gama, João Alves, Carlos Perdigão, Ricardo Pires, Pedro Patrão, Bruno Malveiro, Filipa Ramalho, Raquel Serra and Ana Ribas for their care and help. Special thanks to Luís Soares, for the patience, availability, revision of the work and all the suggestions.

Lisboa, 19 de Julho de 2010
Rui Pedro dos Santos Correia

# Resumo

Num mundo onde a globalização de estudantes e trabalhadores está a aumentar, a área de Ensino da Língua Assistido por Computador vem ajudar a estreitar as distâncias impostas pelos diferentes idiomas. O sistema REAP (originalmente desenvolvido na Universidade de Carnegie Mellon), inserido nessa mesma área de investigação, surge no contexto social actual com o intuito de suprimir a necessidade de adquirir proficiência num dado idioma de uma forma rápida e apelativa para o utilizador.

O REAP.PT (a versão portuguesa do sistema) encontra-se actualmente em fase de adaptação para o Português Europeu. Mais especificamente, o presente trabalho vem dar continuidade a uma primeira tarefa de adaptação, centrando-se na tarefa de geração automática de perguntas de vocabulário. Para alcançar esse objectivo existem várias tarefas que foram desenvolvidas no âmbito da presente tese:

- percepção de que tipos de perguntas são pertinentes e capazes de ser geradas automaticamente e que formas de apresentação podem estas assumir quando apresentadas ao aluno;
- integração de recursos capazes de suportar a geração de perguntas;
- modificações na arquitectura do REAP.PT capazes de representar, de uma forma coerente e conexa, os conceitos envolvidos no sistema e consequente inclusão de dados relevantes para representar esses conceitos (como por exemplo, distinção entre o *lema* de uma palavra e inflexões correspondentes e a classificação das palavras por nível);
- desenvolvimento de estratégias inteligentes para gerar distractores (“respostas erradas”) no caso de se tratar de perguntas de escolha múltipla.

Para além destas tarefas directamente relacionadas com o tema principal da presente tese, houve também o esforço de alargar e melhorar da funcionalidade geral do REAP.PT. Algumas das tarefas desenvolvidas neste âmbito foram:

- adaptação das actuais arquitecturas utilizadas para lidar com grandes quantidades de informação, a fim de processar um novo corpus de documentos, *ClueWeb09*;
- desenvolvimento e adaptação de uma interface para o professor;
- melhorias de apresentação no módulo de compreensão oral do sistema, mais precisamente nas transcrições automáticas de noticiários televisivos.

# Abstract

In a world where globalization of students and workers is increasing, the Computer Assisted Language Learning research area comes as an aid to tighten up the gap imposed by language. The REAP system (developed at Carnegie Mellon University), as an example of a CALL tutoring system, arises in the present social context with the goal of suppressing the need of acquiring a certain degree of proficiency in a language in a fast and appealing way for the user.

REAP.PT (the Portuguese version of the original system) is currently in a porting stage to European Portuguese. In particular, the present work is the follow up of a first porting task, this time focusing its attention on the topic of automatic generation of vocabulary questions. To successfully accomplish this goal there are several tasks that were developed in the scope of this thesis:

- perception of what types of questions are relevant and capable of being automatically generated and what forms can those questions assume when presented to the student;
- integration of resources that are able to support question generation;
- modifications in the REAP.PT architecture that allow a coherent and connected representation of the concepts involved in the system, and inclusion of relevant information to represent those concepts (for instance, distinction of word's *lemmas* and correspondent inflections and classification of words per level);
- development of intelligent strategies to generate distractors ("wrong answers") in what concerns multiple-choice questions.

Apart from these tasks directly related with the main research topic of the present thesis, there was also an effort in terms of the extension and improvement of REAP.PT's general functionality. Some of the tasks developed on this topic were:

- adaptation of the current architectures used to deal with large amounts of data, in order to process a new corpus of documents, *ClueWeb09*;
- development and adaptation of a teacher interface;
- presentation improvements of the oral comprehension module of the system, more precisely in the Broadcast News transcripts.

# Palavras Chave
# Keywords

## *Palavras Chave*

Geração Automática de Distractores

Perguntas de Vocabulário

Ensino da Língua Assistido por Computador

Português

## *Keywords*

Automatic Distractors Generation

Vocabulary Questions

Computer Assisted Language Learning

Portuguese

# List of Abbreviations

AFS Andrew File System is a distributed networked file system which uses a set of trusted servers to present a homogeneous, location-transparent file name space to all the client workstations, having its primary use in distributed computing. It was developed at Carnegie Mellon University.

ARC Internet ARChive is a lossless data compression and archival format by System Enhancement Associates (SEA), also used as a file extension for several different file types that have in common that they are some kind of archive files.

ASR Automatic Speech Recognition system is a system that converts spoken words to text.

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

BN Broadcast News.

CALL Computer Assisted Language Learning, or Computer Aided Language Learning, is an approach to teaching and learning in which the computer and computer-based resources are used to provide an interactive system and materials for language assessment.

CSS Cascading Style Sheets is a style sheet language used to describe the presentation semantics (the look and formatting) of a document written in a markup language.

GB Gigabyte is a multiple of the unit byte for digital information storage and is equal to $10^9$ (1 billion short scale) bytes.

HTC High Throughput Computing is a computer science term to describe the use of many computing resources over long periods of time to accomplish a computational task.

HTML HyperText Markup Language is the predominant markup language for web pages, providing means to create structured documents over the Web.

INESC-ID Institute for Systems and Computer Engineering: Research and Development 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.

L1 First Language, also designated by Native Language or Arterial Language, is the language a person has learned from birth or speaks the best.

L2 Second Language, also designated by Foreign Language, is any language a person knows in addition to his/hers Native Language.

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

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

POS Part-Of-Speech is the role that a word or phrase plays in a sentence.

RCS Revision Control Systems allow the management of changes to documents, programs, and other information stored as computer files.

RDF Resource Description Framework is a metadata model standard proposed by the World Wide Web consortium (W3C).

REAP REAder-specific Practice is a tutoring system developed at Language Technologies Institute (LTI) of Carnegie Mellon University aimed at teaching vocabulary to non-native speakers through reading practice.

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

SVN SubVersioN is a Revision Control System founded and sponsored in 2000 by CollabNet Inc.

TB Terabyte is a multiple of the unit byte for digital information storage and is equal to $10^{12}$ (1 trillion short scale) bytes.

URI Uniform Resource Identifier is a string of characters used to identify a name or a resource on the Internet.

URL Uniform Resource Locator is a Uniform Resource Identifier (URI) that specifies where an identified resource is available and the mechanism for retrieving it.

W3C World Wide Web Consortium is an international community that develops standards to ensure the long-term growth of the Web.

WARC Web ARChive is a revision of the Internet Archive's ARC File Format that better supports the harvesting, access, and exchange needs of archiving organizations, storing secondary content, such as metadata.

WSD Word-Sense Desambiguation is a task of Natural Language Processing that deals with the identification of the meaning of a word that is used in a sentence, when this word has multiple meanings.

WWW World Wide Web, or the Web, is a system of interlinked hypertext documents accessed via the Internet, described as the universe of network-accessible information.

XML Extensible Markup Language is a set of rules for encoding documents in machine-readable form, defined in the XML 1.0 Specification produced by the World Wide Web consortium.

# Contents

# List of Figures

# List of Tables

# 1 Introduction

Communication is at the core of modern societies and, as globalization is becoming increasingly part of the current social context, the need to easily communicate in several languages is growing. On the other hand, the time factor calls for new methods of learning; ideally fast, dynamic and appealing ways of acquiring the ability to understand and speak a foreign language.

Technology has a major role in conceiving systems that fulfill such requirements because of its extremely wide scope and constant evolution.

These facts were the starting point for CALL (Computer Assisted Language Learning) in general and REAP.PT (REAder-specific Practice PorTuguese) in particular. As the name states, REAP.PT is based on the importance of reading activity as a way to become proficient in a new language. Therefore, from the standpoint of the student, the learning method can be summarized in two main phases: text reading and question answering. The text reading phase, along with the development of a stable version of the system, was the main focus of Luís Marujo's Master Thesis [1]. The latter phase, the one involving question generation, is the motivation for this document.

One can imagine two very different ways to accomplish this task. The first one is to have manually generated questions, a time-consuming task, which is not very flexible. The second way is to delegate to the machine the responsibility to generate such questions, maintaining quality, adding speed and the aforementioned desirable flexibility factor.

Although the fact that REAP.PT is the Portuguese version for REAP[1] (Carnegie Mellon University version) and the existence of a substantial quantity of studies in the area of automatic question generation, the task of porting the system to a language as morphologically rich and topologically different from English as Portuguese is not trivial. The singularities of European Portuguese (for example, gender distinction in adjectives or the great number of tenses and inflections for a verb) make this difficult task even more challenging.

[1] http://reap.cs.cmu.edu/ (last visited in July 2010)

## 1.1 Goals

The focus of the present work is, as the name states, to endow REAP.PT with automatic question generation mechanisms. A team of engineers (from L$^2$F) and linguists (from University of Algarve and University of Lisboa) have been working together to develop the system, trying to meet the goals of REAP.PT, as a project of the Carnegie Mellon University-Portugal dual PhD program.

The task of question generation demands a significant amount of lexical, syntactic and semantic resources that, until now, were not present in REAP.PT. For this reason, one of the main goals of the present work is to collect, study and integrate quality resources that can support the question generation task.

With the appropriate resources, one can focus in the generation task itself. It is important to notice that, since REAP.PT aims to teach vocabulary, the questions that should be developed are vocabulary questions only (syntactic and semantic-related questions being a whole new area of research).

Apart from questions that can be directly extracted from the aforementioned resources (that intrinsically represent direct relations between words, as synonymy), there are a specific group of questions that will have a special focus in this work: *cloze* questions. Also known as *fill-in-the-blank* questions, they are a great challenge, demanding several resources. A thorough study was done focusing on the generation of the distractors only. In a multiple-choice question, distractors are the "incorrect" choices that are presented to the student, that try in some way to distract him/her from the correct choice. The main goal is to be able to generate distractors that do not fit in the blank space of the sentence, but, at the same time, present some difficulties to the student (i.e., they are not obviously wrong choices) and test particular aspects and idiosyncrasies of the Portuguese language.

In addition to these resource-oriented tasks, this second porting job will also extend the general functionality of the REAP.PT system; some of those functionalities could only be executed given the work accomplished in the scope of the question generation task.

A paper on the research task of generating distractors for the *cloze* questions was accepted at INTERSPEECH 2010 Satellite Workshop[2], "Second Language Studies: Acquisition, Learning, Education and Technology", co-organized by AESOP, SLaTE, NICT and LASS:

[2] Rui Correia, Jorge Baptista, Nuno Mamede, Isabel Trancoso, and Maxine Eskenazi. Automatic Generation of Cloze Question Distractors. In *Second Language Studies: Acquisition, Learning, Education and Technology*, Tokyo, Japan, September 2010.

[2] http://www.gavo.t.u-tokyo.ac.jp/L2WS2010/ (last visited in July 2010)

## 1.2 Structure of this Document

The present thesis consists of 5 chapters, structured as follows:

- Chapter 2 starts by introducing CALL and REAP concepts, explaining briefly their appearance, main features and state of the art. Regarding REAP.PT, a general perspective of its architecture is presented in this chapter. Then it focuses on several types of questions that can be used to assess vocabulary knowledge and how those types of questions can be presented to the students (what forms they can assume).
- Chapter 3 describes the main necessary modifications to the REAP.PT architecture to support the task of question generation. The description, inclusion and study of three lexical resources for Portuguese is presented here as a starting point for the remaining tasks. This chapter describes the work developed to generate questions and distractors, presenting an experiment over the distractors generation topic, and ends with an evaluation of the obtained results for that same experiment.
- In Chapter 4 the focus is on some side work that was done on the system that is either related with the main research-topic or was considered as a high-priority task. These tasks include the integration of a new set of documents, the development of new features for the teachers, and improvements on the oral comprehension module.
- Finally, Chapter 5 presents the conclusions and future work.

# 2 State of the Art

This chapter is divided into 3 main sections: the first part focuses on the concept of CALL, defining it and stating its main goals; the second part presents REAP, as a tutoring system of the CALL area, and describes the underlying architecture of REAP.PT; finally, a third section is devoted to a thorough study on what types of questions, presentation methods and main techniques are the state of the art on automatic question generation.

## 2.1 CALL

Computer Assisted Language Learning (from now on referred to as CALL) is a concept that was born in the 1950's alongside the spreading notion of interactivity provided by the computer science field.

Egbert and Petrie [3] provide a succinct definition for CALL, stating that "CALL means learners learning language in any context with, through, and around computer technologies". Intrinsically, this definition leads to the establishment of a strong relationship between learning and technology. So, as technology started to meet the requirements for the implementation of CALL's ideas and evolving towards new ways of interaction and mass resource availability, a large amount of CALL-related software began to be developed and used. From the 1960's up until now, one can find software covering a diverse variety of aspects related with language learning, as well as several distinct methods and data to support that task.

The arrival of the Internet and the World Wide Web as a common and widespread resource can be identified as a major milestone regarding the CALL subject. This fact, along with the possibility to exchange knowledge in an easier way (common to all fields of study), brought to CALL a channel to deploy the projects themselves and make resources available on a much wider scale. Facing the nature of CALL's projects and their respective goals, such technology extended the scope and brought up new questions and solutions.

On the other hand, this milestone also has its drawbacks. The notion of a static, official and unquestionable knowledge source (whether it be a lexical resource or resources from another field) all but disappeared or at least was given less and less importance. The WWW created an enormous cloud of materials and knowledge, and more important than that, it created a dynamic resource, where contributions could now be made all around the world in a simple and rapid way. The main problem with this new vision is that any piece of software that wants to take advantage of the Internet should be conceived itself as a dynamic system, able to select **updated** information but, much more difficult and important, to select **correct** information.

In the present days, as research provides better foundations and approaches, these previously mentioned issues are being overcome. The area of CALL is currently in a state of development and maturity allowing for a real integration of such dynamic systems in real life as a solution and as a tool, and not just as a problem or an object of study.

## 2.2 REAP.PT

REAder-specific Practice PorTuguese (from now on REAP.PT) is a software that proposes an implementation of the ideas of CALL.

The concept behind REAP [4] (the original English version of the software), as described by Collins-Thompson and Callan [5], is that it "is intended to advance the state of the art in information retrieval, as well as research in reading comprehension, by bringing together practical user models of student interests, vocabulary knowledge and growth, and other aspects of reading, with interesting material from large, open collections like the World Wide Web."

As already stated, REAP's teaching strategy has its roots in text reading and question answering tasks. Considering the definition in the preceding paragraph, one should realize that REAP is part of the dynamic systems in the area of CALL that were mentioned in the previous section. Indeed, the two main activities of REAP (reading and answering) are constrained to the needs of **each** student. The ambition behind REAP is in fact to present reading materials that are considered interesting for the student (based on user models) and guide him/her through vocabulary learning.

It is important to understand the general workings of the system, the backbone that enables REAP to accomplish its goals. Having a whole language as an object of study, the system has to define and limit the teaching process somehow. This is done using a list of words that one wants the student to learn about. For the Portuguese version, such a resource is materialized by the Portuguese Academic Word List (P-AWL) [6]. The current documentation of P-AWL [7] presents this list as "a careful selection of common words that may constitute a valid tool for assessment of language proficiency at university level, irrespective of scientific or technical domain. One can view P-AWL as a landmark, useful to measure the students' progress on their learning process and language proficiency." The current version is composed by 2,019 different *lemmas* divided by different part-of-speech (POS). Table 2.1 presents the distribution of P-AWL's *lemmas* among the POS categories.

| POS | Number of *lemmas* | % |
|---|---:|---:|
| Noun | 877 | 43.44 |
| Adjective | 480 | 23.77 |
| Verb | 440 | 21.79 |
| Adverb | 216 | 10.70 |
| Conjunction | 4 | 0.20 |
| Preposition | 2 | 0.10 |
| Total | 2,019 | |

Table 2.1: Distribution of P-AWL's *lemmas* along the POS categories.

Apart from the list of words that the system tries to teach, there are two more concepts one needs to focus on, in order to understand how REAP.PT works: *level* and *interests*. It is crucial to identify the level in which the student stands on Portuguese language comprehension in order to successfully guide him/her through the learning process. Without this concept, it would not be possible to present the user with appropriate materials for his/her knowledge stage and, therefore, the whole process of learning would not make sense. This *level* is calculated in the first contact between the user and the system and it is supposed to evolve and adapt according to the clues the student gives to the system – for instance, engagement in reading sessions, assessments or dictionary lookup activities. The concept of *interests* is also a major point regarding the REAP goals – providing user oriented learning processes. *Interests* are recorded, for each student, through the presentation of a survey when the student first logs in to the system. Table 2.2 provides information about what categories are presented in the mentioned survey[3]. With this information, when presenting the set of readings from which the student may choose, the system can prioritize those that seem to match his/hers preferences.

| Category | Example Keywords |
|---|---|
| Movies and Theater | *Movies, theater* |
| Music | *Concerts, instruments, composers* |
| Visual Arts | *Painting, ceramics, photography* |
| Computers and Technology | *Software, electronics* |
| Business | *Jobs, economy, companies* |
| Mathematics, Physics and Chemistry | *Mathematics, physics, chemistry, astrology* |
| Biology and Environment | *Biology, environment issues* |
| Social Sciences | *Psychology, languages* |
| Health and Medicine | *Medicine, diseases, hospitals* |
| Fitness and Nutrition | *Exercise, nutrition, healthy lifestyles* |
| Religion | *World religions* |
| Politics | *Governments, campaigns, political issues* |
| Law and Crime | *Legislation, organized crime* |
| History | *World history* |
| Sports | *Cycling, swimming, soccer* |
| Outdoor Activities | *Surfing, climbing, skiing* |

Table 2.2: Categories presented in the *interests*' survey and examples of keywords.

[3] This set of categories was directly imported from the English version and does not match exactly the topics that the document *classifier* supports, since the classifier is more suited to Broadcast News stories [8].

Along with a specific *level* for each student, a set of the words from P-AWL is assigned to him/her. This set represents the words that he/she should learn, the so-called *focus words*. The way REAP chooses to teach this specific set of words is by providing documents, highlighting the *focus words* and presenting them in an explicit and appealing context for that particular student, given his/her *interests*. Each reading is then followed by an assessment that tests if the learning of the new vocabulary was successful.

REAP began being developed at Carnegie Mellon University, Pittsburgh, PA, and was intended to help teach English as a Second Language. The collaboration between Portuguese and American universities brought the system to the Portuguese researchers table, establishing it as a possible substantial gain in our society.

### 2.2.1 General Architecture

To have a better notion of the way the system works and how the features that will be proposed can be integrated in REAP.PT, it is necessary to understand the system's architecture. Figure 2.1 shows REAP.PT's architecture, as described in Luís Marujo's Master Thesis [1].

Using an appropriate Web browser, users can interact with the REAP.PT system, through the World Wide Web, using the *Web Interface* component. This interface is also responsible for exchanging information between two other modules: the oral comprehension module and the database.

The oral comprehension module, represented in Figure 2.1 by the logo of DIXI software [9], has been put in place because of the known difficulty in understanding spoken European Portuguese. This new feature, an important and innovative contribution to the REAP project, first integrated in the REAP.PT version, aims to provide audio playing options for the readings (text-to-speech) and other materials, such as multimedia documents.

The database is the core of the system. In a brief description, it contains user information (*interests*, *level*, *focus words*, etc), maintains logs of the activities engaged by the users (readings, assessments, word dictionary lookups, etc), stores the words from P-AWL and the questions to be asked and, finally, stores information about the documents (readings) that can be presented.

This latter function (document storage) is related to the not yet described elements in Figure 2.1. These have to do with filtering and classification of the documents that will be stored and presented. WPT05 Web corpus[4] "is a collection of 10 million documents from the Portuguese web obtained by the crawler of the Tumba! search engine[5]" and constitutes the original document source. As one might expect, not all the documents are appropriate for presentation to the student, whether by containing inappropriate/irrelevant content or invalid format. For that reason, the WPT05 corpus is submitted to

[4] http://xldb.di.fc.ul.pt/wiki/WPT_05_in_English (last visited in December 2009)

[5] http://xldb.di.fc.ul.pt/wiki/Tumba! (last visited in December 2009)

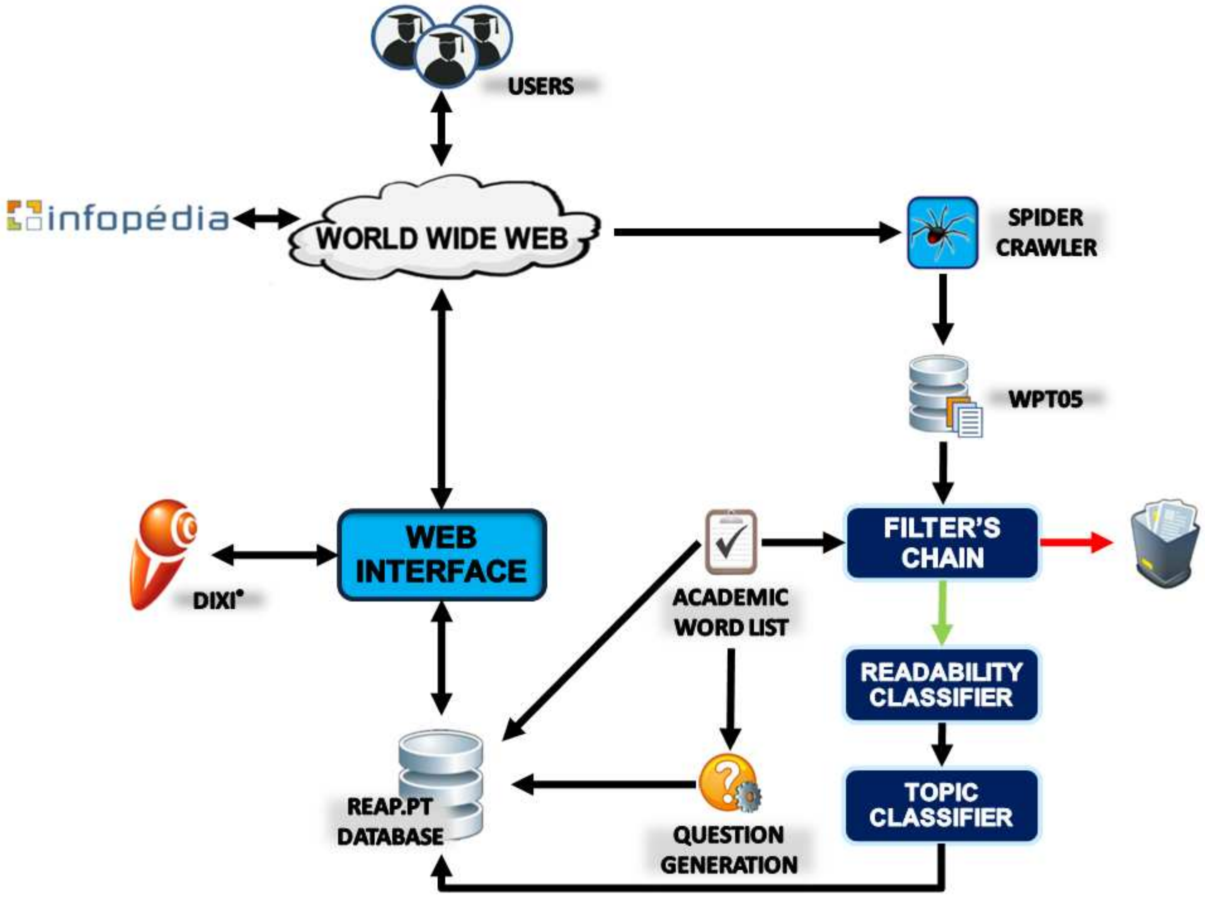


Figure 2.1: Architecture of REAP.PT system.

a chain of filters that excludes documents that are not in HTML format, are too short (less than 300 words), contain profanity words, do not include words from the P-AWL, or are just lists of words. After this filtering task, documents are classified by readability level and topic (allowing future crossing with user's *level* and *interests*), and are finally stored in the database. This specific part of REAP.PT's architecture had a special focus during the present thesis and will be discussed and detailed later in Section 4.1.

## 2.3 Automatic Question Generation

Although knowing that crafting exercise questions is a time-consuming task one might be tempted to believe that it is a task that only needs to be done once. However, that is not true!

REAP.PT is a system that has several target users. General Portuguese second language (L2) learners are the main target, but one may want to apply the system in different contexts. For example, one may use it to teach vocabulary about medicine or economy, allowing students to learn words from a specific domain. Another reason why manually generated questions are not quite a good solution is the inevitable evolution of the language and the changes of the evaluation parameters. In fact, in today's juncture, Portuguese speakers are dealing with the standardization between European and Brazilian Portuguese (a new agreement on uniform spelling has been adopted [10]). One final reason relates to one of the main goals of REAP – providing recent and relevant readings for today's reality. This fact implies that the set of readings is also a dynamic resource, which is supposed to be continuously growing and being updated. These assertions are sufficient to conclude that neither P-AWL, specifically, nor software with REAP's goals, in general, should actually rely on static approaches of a language.

The need for a dynamic system brings two major drawbacks. In the first place, relying on automatic methods makes the control task harder. When using automatic methods one needs to have an extremely high awareness of what is being done in order to control and manipulate it, according to one's intention. On REAP.PT in particular, this difficulty arises in the assessment phase, where articulation between the student's knowledge, the different types of questions and the questions themselves must be performed. The second drawback comes from the need to guarantee the quality of the questions, which is, in fact, an extremely important requirement for this kind of application. To fulfill this requirement, Natural Language Processing tools must be applied.

One of the major points regarding questions' quality is context. This problem is common to most question types that will be presented later (excluding *reading-check* questions). In order to present the student with good quality questions, it is important to know which sense of the word was used in the presented text. If the text is automatically part-of-speech (POS) annotated, and if for a particular word and part-of-speech there is only one sense, one knows exactly what the meaning is. If, on the other hand, for the same POS there are at least two meanings, then one can use the context in which the word was presented to try and extract the exact sense, matching this context with the definition of the word in a resource with semantic description (with example sentences, for each sense, for example). Still on the problem of two senses for the pair word-POS, a much simpler solution is to select the most frequent sense, if such frequency data is available. In conclusion, without context one can not apply the first technique (cross contexts), and without part-of-speech a high error rate is expectable if one just chooses the most frequent sense.

Let us illustrate this problem with a practical example. According to the Longman Active Study Dictionary [11]:

bank$^1$ n **1** the company or place where you can borrow money, save money etc: *I went to the bank at lunchtime to pay in my salary*. **2** land along the side of a river or lake: *trees lining the river bank* **3** a place where a type of thing is stored until someone needs it: *a blood bank* **4** a large pile of snow, sand etc.

bank$^2$ v **1** to put or keep money in a bank **2** to make a plane slope to one side when it is turning: *The plane banked and turned toward Honolulu.*

This entry transcription really points out the senses problem. The word *bank* can be either a noun or a verb. And even if one knows the part-of-speech for the word (whether it is a noun or a verb), the correct sense of the word can not be extracted in a trivial way since it has different meanings in either case.

The previous example denounces another major problem regarding automatic question generation – word derivation. Does the fact that the student knows the meaning of the word *bank* imply that he/she also knows the word *banker*? And what about the reverse situation? Therefore, it is important to keep track of this type of interaction between the student and the system and try to find which questions can be asked, i.e., if the testee reads *banker* in the text, is it legitimate that he/she would be asked about the word *bank* and vice-versa? In either situation, these relations between words and their possible variations should be stored in order to allow a reflection on this subject. These issues represent a challenge and should be considered in the architecture of the system. In fact, they depend on the type of questions that are chosen to be presented as well as on the resources that are available.

In terms of question typology, to define and distinguish them, one needs a formal vocabulary. For this purpose, one should focus on the notation and definition of the constituents of a question.

The ***instruction***'s role is to inform the student what he/she is supposed to do. It is the first information the student is presented to, working as an introduction to the task that is being assigned to him/her, and for this reason should be direct and clear. An *instruction* will look like this: "Choose the word that best completes the phrase below:" or "What is the antonym of *happy*?"

Another crucial element regarding question composition is the ***stem***. When, for a specific question type, the student is asked to complete a particular idea with a missing word, the *stem* is the embodiment of that same idea – it is represented by a sentence with a blank space where the missing correct word should be. The *stem* is most used when testing the student's ability to choose a correct word for a given, well defined context. While presenting the several question types, this element will be discussed in a more precise way.

Another element regarding question constitution is the set of ***answer choices***. As the name states, this set is composed by words, or other elements that are the possible answers to the question being asked (for example, definitions), from which the student has to choose one. One of the *answer choices* is the ***correct answer*** and the remainders are called ***distractors***. The *answer choices* set is used when the question is a *closed* one, which is defined exactly by the fact that several choices are provided. If there is no such set, one says that the question is an *open* one.

The remainder of this section will introduce a set of relevant question types, ways to present them and talk about *distractors* generation.

### 2.3.1 Typology of Vocabulary Questions

Focusing on the major goal of REAP.PT, vocabulary learning, special attention must be given to the choice of the questions the students should be asked. All questions ought to have vocabulary assessment as a focus, but just as important is the ability of the system to track the students' knowledge level. Each type of question, whether rightly or wrongly answered, must give an indication to the system about the *level* the testee is in.

Diversity in the type of questions, apart from enabling a less fastidious learning method, allows a response to the classification of the student's *level* issue just brought up. Perceiving which type of questions measures which phase of word knowledge can help the establishment of an order of appearance of the questions in the assessment phase, enabling efficient control and guidance of the student through this phase.

Stahl [12] categorized word knowledge in three main stages that can be used to assign a state of knowledge of the word for a given student:

- *association processing* – ensures that the student is able to associate the new word meaning with previous familiar concepts;
- *comprehension processing* – has to do with being able to understand the meaning of the word in a given context;
- *generation processing* – requires the ability to use the word in several contexts showing a deep knowledge of the word's meaning.

Having recognized that vocabulary acquisition is a phased process – and a phased and organized assessment stage can suit this process – one will now present six types of questions: reading-check questions, definition questions, synonym/antonym questions, hypernym/hyponym questions, *cloze* questions and open *cloze* questions. Each one will be addressed in terms of constitution, stage of vocabulary knowledge and specific features.

#### 2.3.1.1 Reading-Check Questions

This type of question is a special one, since it does not involve complex lexical resources and does not test vocabulary acquisition. Reading-check questions, as the name states, aim to check if the testee demonstrates having actually read the text.

Figure 2.2 represents an example of this type of question. The student is asked to choose which set of words is fully comprised by words that appeared in the text he/she just read. Here, each *answer choice* is a set of words: for the correct answer, all the words appeared in the text; the *distractors* contain a subset of the words that are present in the correct answer, and are filled with other words that did not appear in the text.

Choose the set of words that appeared in the document you just read.

❏ power – house – mail – horse – electricity – care – supply – top

❏ torn – mail – environment – electricity – shore – horse – wind – vague

❏ electricity – salesman – suit – attention – supply – food – winter – sprint

❏ care – abuse – house – vending – nature – young – vague – salesman

Figure 2.2: Example of a reading-check question.

For the already mentioned reasons (no dependence of complex lexical resources and use of words extracted directly from the reading), reading-check questions can be completely automated in an error free way.

However, there are some common sense guidelines that can be followed in order to improve the effectiveness of this type of question. The words that comprise the correct answer should be expressive in such a way as to reflect the subject of the text. On the other hand, they should not be so much related between them as to give away the answer too easily. Words that the system is trying to teach (*focus words*) should be avoided as well, because the student may already be accustomed to them.

Feeney and Heilman [13] focused on this subject. They introduce the concept of *salience* of a given word, $w_i$, in a given text as the frequency of the word $w_i$ in the text minus the frequency of the word in the language (English, for that case). The division of this measure by the frequency of the word in the language has the intent to give low values to words with high frequency rate (in Portuguese, for example, words like “os” or “e”, the same as “the” and “and”). The equation that computes the *salience* of a given word, $w_i$, in a given text *j*, is defined in Equation 2.1, where *V* is the number of possible words and *freq($w_i$)* is the frequency of a given word, w$_i$ in the language. This last measure (frequency of a word in the language) is defined in Equation 2.2, where *D* is the set of documents that are being considered for the word count.

$$S_j(w_i) = \frac{\frac{count_j(w_i)}{\sum_{k=1}^{V} count_j(w_k)} - freq(w_i)}{freq(w_i)} \tag{2.1}$$

$$freq(w_i) = \frac{\sum_{m=1}^{D} count_m(w_i)}{\sum_{m=1}^{D} \sum_{k=1}^{V} count_m(w_k)} \tag{2.2}$$

REAP.PT should also be concerned with rarer words (like proper names) and discard them from the sets of *answer choices*.

Regarding *distractors*, they may be obtained in a random way or, for an increased difficulty, with synonyms or words related to the text context.

Wrong answers to these questions may be a sign that the student has not read or understood the text and this information can be retained as experimental data, used to suggest to the student that he/she should re-read the text or even tell the system that the texts that are being assigned for this particular student are too difficult for him/her.

This type of question is merely a test on the attention degree of the student and does not relate with word meaning knowledge, so the Stahl theory does not apply.

Again, Fenney and Heilman [13] concluded after evaluation that "reading-check questions seem to measure a construct that facilitates vocabulary acquisition while reading practice texts. (...) in general, students who correctly answered reading-check questions also learned more vocabulary." This result suggests that this type of questions might be not only a measure of attention and engagement, but may also help the student to acquire vocabulary, by relating the concepts while trying to understand the global idea of the text.

#### 2.3.1.2 Definition Questions

As the name states, this type of question consists of asking the student to find the correct word-definition pair. Figure 2.3 illustrates an example of a definition question.

Which word does the following definition best describe?

**"Sweet food with high calorie and preservative properties."**

□ salt

□ energy

■ sugar

□ oil

Figure 2.3: Example of a definition question.

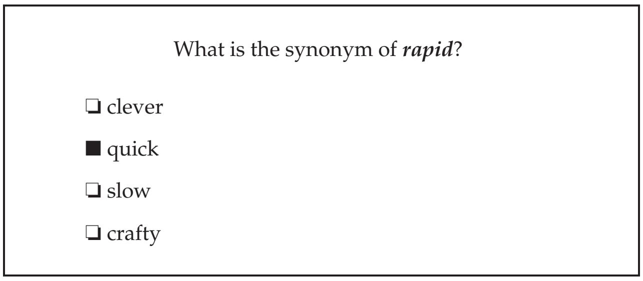
As might be expected, these questions can only be generated based on the existence of a reliable resource for the definition texts. One can conceive two versions of this type of questions: given a definition and several words as *answer choices*, the student has to find the correspondence; or the other way around, given a word and several definitions as *answer choices* the system asks the student to find the perfect match. In addition, it should be noticed that the definition text has to be such that the word that is being tested (or any morphological variation) does not appear in the definitions.

The complexity of such questions relies on the fact that the definition text itself must be related with the sense of the word that appeared in the text - the problem of several senses for the same word previously mentioned.

Analyzing Stahl's theory, if the student correctly answered a definition question, the first stage of word knowledge is accomplished, since the testee is able to identify what the word can and can not mean for, at least, one particular context.

#### 2.3.1.3 Synonym/Antonym Questions

As the previous type of questions, producing synonym/antonym questions is a resource-dependent task, meaning that it can only be achieved having access to a good resource of word relations. Figure 2.4 presents an example of a synonym question.



What is the synonym of ***rapid***?

❑ clever

■ quick

❑ slow

❑ crafty

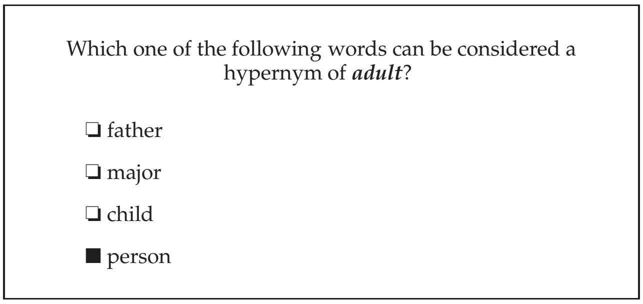
Figure 2.4: Example of a synonym question.

Here one can use the direct synonyms of the word or even synonyms' synonyms. As Brown, Frishkoff and Eskenazi [14] advise, the synonyms that may be considered should be a single word and should not be just a morphological variant of the original word. If more than one word has these properties then the most frequent word should be chosen. This choice can make the question easier, so one may choose the synonym to use according to the testee *level*. The same can be applied to antonym questions, using both the direct antonyms and the antonyms of the synonyms of the *focus word*.

Here, Stahl's second level is met in the way that the student is able to relate two words, recognizing a common/opposite meaning between them.

#### 2.3.1.4 Hypernym/Hyponym Questions

This type of question uses the same strategies as the previous one, thus being a very resource-dependent task. Figure 2.5 illustrates an example of a hypernym question.



Which one of the following words can be considered a hypernym of ***adult***?

□ father

□ major

□ child

■ person

Figure 2.5: Example of a hypernym question.

Hypernymy is a term used to refer to a relation between words where the words involved share a class to subclass relation. Hyponymy represents the inverse relation – subclass to class [15].

Brown, Frishkoff and Eskenazi [14] concluded on their evaluation of the same subject for English that they were not able to generate this type of question for adjectives and hence, for that particular POS, this type of question was excluded. Seeing that, in REAP.PT, adjectives represent 23.77% of the P-AWL (Table 2.1), depending on the coverage of the resources for the remainder part-of-speech, this type of question can be ignored (if there is low coverage) or developed.

For the same reasons as the previous question type, hypernym/hyponym questions can guarantee Stahl's second level.

#### 2.3.1.5 *Cloze* Questions

The family of questions denoted as *cloze* has a unique quality compared to the rest of the types of questions already addressed. This singularity has to do with the fact that the *focus word* is being tested in a specific context.

The word *cloze* was adapted from the word *closure* related with the concept of "Law of Closure", in psychology [16]. This law states that the human brain, in order to increase regularity, is able to conclude information that does not exist and, thus, could not be perceived. Figure 2.6 presents a visual approach of this law, which shows that human brain automatically recognizes a complete circumference when there are missing parts.

Migrating this visual *closure* concept to the vocabulary learning perspective shows the essence of *cloze* questions – to be able to find the word, among the set of *answer choices*, that has a best fit in a specific

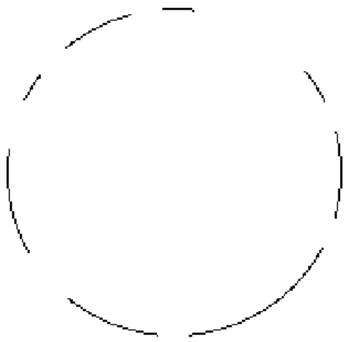


Figure 2.6: Example of the “Law of Closure”.

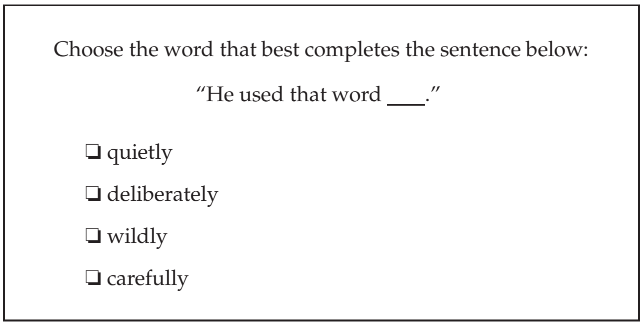
*stem*.

This type of question represents two challenges for a complete automation: the conception of the *stem* and the choice of the *distractors*. Actually, to successfully create a *cloze* question, *stem* and *distractors* have to be in accordance, so that there is no doubt about what word would fit in the blank, i.e., amongst the *answer choices* there is only one possible solution. So, for this task a new kind of resource is required – a resource that has example sentences for each word. Ideally, this resource would be specifically oriented to these sentences, assigning to the word an example sentence that illustrates its usage in a context where no other could be. If no such resource is available, one should notice that any collection of Portuguese texts, provided that it has a sufficient amount of texts, is able to deliver example sentences for *cloze* questions but, since it is not created for this particular goal, it would call for additional work of quality assurance.

It is clear that the sentences that should be used to constitute the *stem* must have a sufficient well-defined context in order not to restrict the number of words that would correctly fill the blank. For this reason, short sentences are not, usually, a good resource to support these questions as one can see in Figure 2.7, adapted from Pino, Heilman and Eskenazi’s work [17]. In view of this, the length of the sentence is a good heuristic to measure the adequacy of a sentence to be used as a *stem* (longer sentences provide more context). Creating a well suited *stem*, with a strong context for the word being tested and, at the same time, with no relation with the *distractors*, is really a hard task. For example, even if the *stem* in the example of Figure 2.7 was longer and established a better context, the fact that the word that is being tested is an adverb increases the complexity, since different adverbs can fit in almost every sentence, if the context is not specifically oriented to test them.

Pino, Heilman and Eskenazi [17] state that, if no such sentence is available, using the word’s set of synonyms (also known as *synset*) to generate the *stem* is an appropriate strategy.

It is important to find methods for measuring *stem* quality and application. Again, Pino, Heilman



Choose the word that best completes the sentence below:

"He used that word ____."

❏ quietly

❏ deliberately

❏ wildly

❏ carefully

Figure 2.7: Example of an ill-defined context in a *cloze* question.

and Eskenazi [17] described a suitable method for this task – a weighted sum which criteria are described below.

- the first criterion is **complexity** – a sentence is more complex than other if it has more clauses. This simple method is based on the premise that a well-suited *stem* is the one that is able to establish a context when testing the word. Usually one clause is needed to set up the context and one clause to test the word;
- the other measure that can be used is to see how well-defined the context is, i.e., if it accepts the word being tested but rejects the presence of any other word. This can be computed using **collocations** [18]. Collocations are sets of words that frequently co-occur within a sentence. With this, using a pre-defined window of adjacent words, one can compare how well a word relates with the rest of the sentence. For this strategy to be applied, one needs a resource that contains such information – words that co-occur. This heuristic signals that the context may be well defined if a word *collocates* frequently with other words in the phrase;
- **grammaticality** is another evaluation method and consists of the submission of each candidate *stem* to a parser to measure the quality of the syntax of the sentence. Here, one has to use the length factor to adjust the scores since shorter sentences typically have higher grammaticality scores;
- finally, length is in fact another important measure to find a good *stem* for the reasons already discussed.

Despite being the most difficult questions to answer, *cloze* questions do not require enough capabilities to satisfy the ultimate level of Sthal's theory. *Cloze* questions guarantee only Stahl's second level.

#### 2.3.1.6 Open *Cloze* Questions

Open *Cloze* questions are a special case of *cloze* questions in which the student has actually to type in the answer in the blank space instead of selecting an answer between a given set of possible ones. Usually, these questions are manually evaluated by the professor, after the assessment.

Obviously, all the strategies to find the appropriate *stem* still apply, but the scope is a little relaxed since there is no need to generate *distractors*. On the other hand, if one wants to test knowledge of a specific word with this type of question, the "set of *distractors*" is the entire set of words that the student knows and finding a good *stem* is even more challenging.

## 2.3.2 Question forms

### 2.3.2.1 Multiple-Choice

This presentation method is the one that has been used in the previous examples of questions of the present document (Figures 2.2 – 2.7). It is composed by the *instruction*, the *stem* (if it is a *cloze* question) and the set of *answer choices*.

One has to pay attention to the wording used for the instructions and to the quantity of *distractors* that are presented. Graesser and Wisher [19] presented directives for these issues on multiple-choice questions, stating that the ideal number of *distractors* is three (3) plus the correct answer, and the *instruction* should be clear and appropriate.

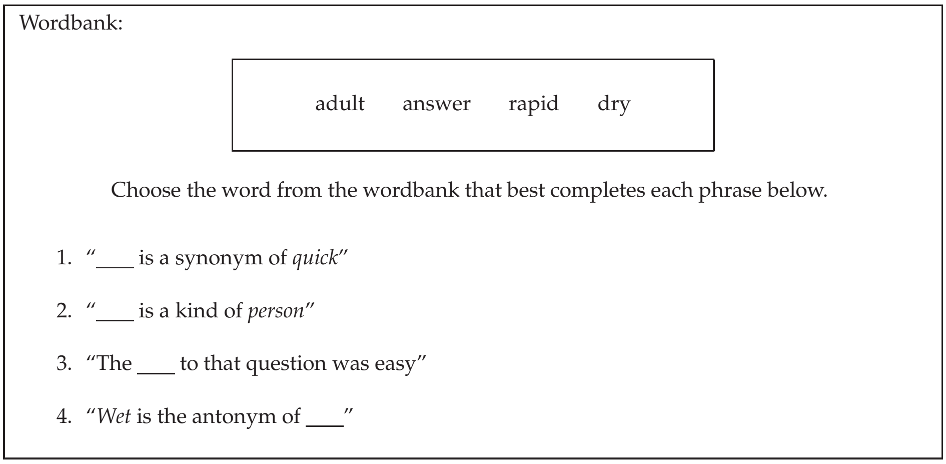
### 2.3.2.2 Wordbank

This is a slightly different way to present the generated questions. The testee is asked to connect the words that he/she sees in a box, the *answer choices*, with the set of *stems* below. The difference here is that all the *answer choices* are at the same time correct answers. There are as many *stems* as the words that form the *answer choices*.

For the questions that typically do not have a *stem* one needs to create this element. In fact, the *stem* can be adapted and used in other types of question than just *cloze* questions. For example, using the instruction "Choose the word that best completes the phrase below:" and the *stem* "____ is a synonym of *quick*" represents the same information as the instruction "Which of the following is a synonym of *quick*?" A possible fictitious example of this type of question is presented in Figure 2.8.

With this kind of question, one can take advantage of the fact that the *distractors* of one question are the correct answers of the rest of the questions, thus eliminating the necessity to generate *distractors*. On the other hand, the fact that *distractors* are not chosen specifically for a particular question, may reduce the difficulty level of the question itself.

The singularities of the wordbank form imply another conclusion: the testee may get a question right just because there is only one word left. From the point of view of the assessment, strictly, this may



Wordbank:

adult answer rapid dry

Choose the word from the wordbank that best completes each phrase below.

1. “___ is a synonym of *quick*”
2. “___ is a kind of *person*”
3. “The ___ to that question was easy”
4. “*Wet* is the antonym of ___”

Figure 2.8: Example of the wordbank form.

be considered a drawback since the student does not have to apply any knowledge to answer correctly to that specific question. But, from the standpoint of the learning process, the student may actually learn the word.

### 2.3.3 Distractors

The use as distractors in *cloze* questions, as opposed to using open questions, meets the requirement of having a system with a certain degree of automation in the learning process, not requiring the manual grading of answers by the teacher. This would have too much impact in the dynamics of the system, slowing down the acquisition of vocabulary by the students. On the other hand, having a proper set of *distractors* associated to each question may also work as a guide, driving the student’s attention into a specific and controlled set of words.

*Distractors* can be extracted from several sources. They can be, for instance, random words, phonetically similar words, antonyms or variations of the correct word. As advised for the English language by Coniam [20], one should use *distractors* within the same POS and frequency rate of the word that is being tested. In Portuguese one may be interested in exploring common mistakes, like spelling errors (such as using “ç” instead of “ss”) or even errors in verb tenses, due to their complex variety (as opposed to English).

For the REAP system, as the student has just read a text with the *focus words*, one should pay attention to the fact that *distractors* may be easy to find if they are generated randomly, by simple noticing that they did not appear in the text. So, it may be important to extract some of them from the text itself. One should also notice that having just read a text, one has conceived a context and *distractors* out of the context (semantically different) can be excluded right away. Finally, excessive repetition of *distractors* may cause the student to get accustomed to them, thus giving away the answer too easily.

*Distractors* can be also scored applying the same score procedure and criteria used to measure the quality of the *stem* by replacing the word that comprises the correct answer with the *distractor* in the *stem*. Choosing *distractors* with the highest scores prevents the ones that are obviously wrong from being presented to the student but can make answering nearly impossible, since more than one possible answer might be appropriate. So, to avoid several possible answers, one should use *distractors* that are semantically "far enough" from the correct word. Patwardhan and Pedersen [21] present a method for computing this semantic similarity. The words that one wants to compare, $w_1$ and $w_2$, are associated with their definition. Then each word, $d$, of that definition is associated with a *first order context vector*, computed by counting the co-occurrences of $d$ with other words in a corpus (a corpus comprised by the definitions of the words, for that particular case). Then, computing the sum of these *context vectors* produces a *second order context vector* that represents the meaning of the word. At last, the dot product of the *second order context vectors* associated with the two words, $w_1$ and $w_2$, gives the semantic similarity between them.

Goodrich [22] presents a way to determine the efficiency of *distractors*. To measure it, two concepts are involved: *potency* and *discrimination*. *Potency* is the percentage of students that make a specific choice. Here, there is a trade-off involved between having nobody choosing a particular *distractor*, indicating that it is not "giving the question a factor of difficulty", and being frequently selected, indicating that it may be a "correct answer to a badly posed question". *Discrimination* has to do with the ability to differentiate students of different levels of proficiency. Goodrich also distinguishes some *distractors* categories. Some of them, like the use of antonyms and false synonyms (words that have close or similar meaning but cannot fit in the context of the stem) are fragile regarding automatic generation, since they are very context dependent, and can easily lead to the generation of correct choices instead of *distractors*. Some of the categories discussed in Goodrich's work were used for the present evaluation, namely random *distractors*, graphemic variation and morphological variation.

Pino and Eskenazi [23] focused on this same subject for the English version of REAP. In their work, each one of the 33 *cloze* questions was assigned to a set of *distractors*, each one belonging to a different category (morphological, orthographic, phonetical, and combinations of orthographic-morphological and phonetical-morphological). This study was aimed at non-native speakers, relating the origin of the student with the category that proved to work better as a *distractor*. In fact, the native language of the student proved to influence the *distractors* choice.

# 3 Automatic Question Generation

This chapter is devoted to the research topic of automatic question generation in the REAP.PT system. It is divided intp 2 main sections.

The first part, Section 3.1, describes the modifications that took part in the REAP.PT architecture in order to be able to generate the proposed type of questions. This encompasses the addition of new resources, new representations of REAP's concepts and articulation between them. This first section ends with a description of the final (and current) architecture of the system, and with an enumeration of the capabilities added to the system with the new setup.

Section 3.2 explains how definition, synonym and hypernym/hyponym questions were generated and integrated in the system. It also focuses on the topic that was most explored during the present thesis – generation of *cloze* question *distractors*. The description of the experimental setup and the correspondent results are presented in this section.

## 3.1 Preparing the System

The REAP.PT architecture, more specifically, the REAP.PT database is now endowed with the necessary capabilities and data for question generation. Therefore, this section presents the contents that were added to the database to support the proposed task. The recently included contents are an extension of the P-AWL, the level of the *focus words* (words that the system is trying to teach), the *stems* for the *cloze* questions and, finally, lexical resources.

### 3.1.1 P-AWL extension

As described in Chapter 2, there is a list of vocabulary (P-AWL) that was specifically developed by a team of linguists in the University of Algarve for REAP.PT project. In the same chapter one pointed out the problem of word inflections, i.e., the variations in gender, number, tense, etc, of the original *lemmas*, highlighting the necessity to keep track of these variations in the documents and questions that are presented to the student. In sum, it is important to define a new concept that represents the word inflections, and assure a relation between a *lemma* and the respective inflections.

In order to populate the database with the inflections of the *lemmas* proposed in P-AWL, one used a file that had a description of each of the inflections for all POS except for verbs (gender and number). For the verbs, a verbal conjugator[6] was adopted. A new table, "WordForm", was created in order to store these inflections. "WordForm" is then related with the table "Word" that stores the *lemmas*. With this new table, it is possible to know the inflections of a given *lemma* and vice-versa, the *lemma* of a given inflection form. "WordForm" table represents the list of *focus words* of the system, i.e., words that can be highlighted in the readings, and upon which questions may be asked.

Extending P-AWL with the inflections[7] of the original *lemmas* left the table "WordForm" with 33,284 entries. One should notice that for a verb there are, in most cases, sixty-eight (68) variations, which totals, only for the verb *lemmas*, 29,337 forms. Table 3.1 presents the distribution of the *lemmas* facing their inflections, where the predominance of the verb forms is clearly visible.

| POS | *Lemmas* (%) | Inflections (%) |
|---|---:|---:|
| Noun | 43.44 | 5.50 |
| Adjective | 23.77 | 5.70 |
| Verb | 21.79 | 88.14 |
| Adverb | 10.70 | 0.64 |
| Conjunction | 0.20 | 0.01 |
| Preposition | 0.10 | 0.01 |

Table 3.1: P-AWL and inflections of the *lemmas*.

Since P-AWL is a dynamic resource that is still currently being developed, it will certainly change during REAP.PT's development and even usage. This automatic mechanism that creates a consistent list of *focus words* will be usefull in face of future modifications.

### 3.1.2 Words' Level

As one might expect, the information about the level of the *focus words* is essential to be able to guide the student along the learning process that REAP.PT proposes. Each student should be assigned a list of words from the level that represents his/her proficiency in the Portuguese language. Ideally, the level of a word would represent the complexity of that same word in the Portuguese language, as it is reflected by the school year this word is introduced and the year at which its mastery is deemed to be achieved. Apart from this fact, as already mentioned, level data is necessary for generating *distractors* and *stems* for the questions (that should have the same level as the word that is being tested). Again, the existence of automatic mechanisms to accomplish this is of major importance, in view of possible future modifications of P-AWL.

[6] developed by Fernando Batista in 2000.

[7] for nouns and adjectives – gender and number variation; for verbs – variation in number, person and gender (when applicable) of the following tenses: present indicative, preterite indicative, imperfect indicative, pluperfect, future indicative, present subjunctive, imperfect subjunctive, future subjunctive, past participle, conditional, infinitive, personal infinitive, imperative and gerund.

To do this classification one adapted the readability classifier developed by Marujo [24] to classify the readability level of a given text. The main idea of the method is to use *unigrams* (words' counts) to measure the probability of a word belonging to a certain grade level.

The core of the algorithm is the construction of *Language Models*. The training and test corpora used to compute the mentioned *Language Models* consist of 47 textbooks and exercise books (made available by the Portuguese publisher Porto Editora[8]), the held-out test set being composed of one book per level. This test set of books was complemented by a set of national exams for the 6th, 9th and 12th levels (5, 7 and 6 exams, respectively). The input for the algorithm is therefore a set of **.txt* files which contain the unigrams counts for each level. Then two hash maps are used:

- *wordCountInLevel* – maps a word with the corresponding probability of the word appearing in a given level. This probability is computed using Equation 3.1. The probability of a given word, $w_i$, being in the level $l_j$ is the proportion between the number of occurrences of $w_i$ in $l_j$ and the total number of words in the level $l_j$ (where *N* is the number of different words in the level);
- *matrixWordFrequencyPerLevel* – maps each grade level with the corresponding *wordCountInLevel* hash table.

$$P(w_i = l_j) = \frac{\#(w_i, l_j)}{\sum_{k=1}^{N} \#(w_k, l_j)} \tag{3.1}$$

With this information, the algorithm executes the final step – builds a hash map that maps words into levels. To accomplish this, each word is first assigned with the first grade level they appeared in. Then, with a window of two grade levels, the algorithm will try to find if the probability of the word in those two following levels is higher and, if it is, it will be assigned to the level that has the greatest probability for that particular word.

The main problem with this approach relies on the classification of very common words that end up associated with higher levels (two levels, maximum) since they are used transversely in all levels. For this reason, one recommends further research on the topic of readability classification, in order to be able to assign low readability levels to very common words (such as function words).

Table 3.2 presents the distribution of the P-AWL words (just *lemmas*) along the several grading levels.

As one can see, there was a considerable amount of words that were not assigned to any level (18%). One should notice that these results are based in the unigram counts of the *lemmas*, i.e., for

[8] http://www.portoeditora.pt/

| Level | % of P-AWL's words |
|---|---:|
| Five | 10.58 |
| Six | 9.93 |
| Seven | 17.58 |
| Eight | 11.17 |
| Nine | 6.46 |
| Ten | 12.51 |
| Eleven | 7.99 |
| Twelve | 5.66 |
| No level assigned | 18.12 |

Table 3.2: Distribution of the P-AWL's words between grading levels considering only the *lemmas*.

example, if the word "cars" exist in the corpus, but the singular "car" does not, the *lemma* "car" will not have a readability level assigned. For that reason, one tried a second approach to classify the words from P-AWL: instead of counting only the original *lemmas*, consider also their inflections. Following the example just described, the counts for the word "cars" were assigned to the respective *lemma*, "car". Table 3.3 presents the distribution of the P-AWL words according to the second approach.

| Level | % of P-AWL's words |
|---|---:|
| Five | 10.95 |
| Six | 11.39 |
| Seven | 21.74 |
| Eight | 10.90 |
| Nine | 6.24 |
| Ten | 11.94 |
| Eleven | 7.78 |
| Twelve | 5.05 |
| No level assigned | 14.01 |

Table 3.3: Distribution of the P-AWL's words between grading levels considering inflections.

As one would expect, apart from the fact that there were more words classified, this new method assigns a lower level to some words that had a higher classification with the preceding approach. The reason for this classification lies in the fact that, with the new method, the probability of a word appearing in a lower level is higher, since one are now considering its inflections.

*Appendix A* provides a complete list of the words that were assigned to each level, according to the second method described.

### 3.1.3 *Cloze* Question Stems

As a first approach to automatic question generation one relied on a set of manually generated stems for the *cloze* questions that is still being developed by our project partners from University of Algarve.

The goal is to select sentences for all the P-AWL words (*lemmas* and most common infections), that are suited to be presented as a *stem* in a *cloze* question. In view of this, each sentence was selected from text and web corpora, according to a predefined set of criteria:

- one should use only full sentences and not fragmentary text;
- titles, captions, and other paratextual elements should not be used;
- sentences should not have definitions or any othe lexicographic context;
- the target word should not be at the beginning nor at the end of the sentence;
- sentences should be short but not too short, between 100 to 200 characters;
- in the case of ambiguous words, different meanings are represented by independent sentence sets;
- sentences should correspond to a "natural" or "characteristic" distribution of the target word;
- sentences should constitute a non ambiguous environment for the correct identification of an ambiguous word;
- sentences should provide a balanced set of each target word, and its most current inflected forms;
- sentences can be shortened or slightly modified in order to make them comply to these criteria.

With the new data of *lemmas* and correspondent inflections in the database it was possible to develop a mechanism to insert the *stems* with a validation step, i.e., being sure that the tuple word/POS/inflectional categories/stem is correctly described. This new mechanism allowed the detection of errors in the files that described the *stems* and even in the P-AWL, which contributed to the improvement of the quality of these resources. The fact that the *stems* are now inserted in the database with a correspondence to the words they aim to test, makes it easier to select an appropriate question when in the assessment phase.

Currently, there are 3,890 stems validated in the database and they are distributed along POS and morphological categories as presented in Table 3.4.

### 3.1.4 Three Portuguese Lexical Resources

As described throughout the present document, lexical resources are the root for the automatic question generation task. When dealing with vocabulary acquisition, it is crucial that the resources that will provide contents to the question generation module fulfill two main requirements: quality and high recall. Without good resources (those without mistakes or flaws), no matter what methods are applied, it is impossible to generate relevant and correct questions. In the same way, if they are not exhaustive and do not cover a major set of P-AWL's words, the generation would be also impossible.

| POS | Inflectional categories | Number of stems |
|---|---|---|
| Adverb | – | 235 |
| Adjective | Male; Singular | 337 |
| | Male; Plural | 153 |
| | Female; Singular | 179 |
| | Female; Plural | 140 |
| Noun | Male; Singular | 495 |
| | Male; Plural | 261 |
| | Female; Singular | 695 |
| | Female; Plural | 318 |
| Verb | Gerund | 53 |
| | Infinitive | 231 |
| | Present; $3^{rd}$ Person; Singular | 262 |
| | Present; $3^{rd}$ Person; Plural | 13 |
| | Imperfect; $3^{rd}$ Person; Singular | 132 |
| | Imperfect; $3^{rd}$ Person; Plural | 1 |
| | Past Perfect; $1^{st}$ Person; Singular | 4 |
| | Past Perfect; $3^{rd}$ Person; Singular | 154 |
| | Past Perfect; $3^{rd}$ Person; Plural | 8 |
| | Past Participle; Male; Singular | 170 |
| | Past Participle; Male; Plural | 15 |
| | Past Participle; Female; Singular | 31 |
| | Past Participle; Female; Plural | 3 |

Table 3.4: Distribution of the stems by POS and inflectional categories.

With this idea in mind, and knowing that there is not a perfect resource, the solution lies in using more than one resource, each one of them covering different aspects of the language. Three resources were chosen: PAPEL, MWN.PT and TemaNet. The following sections, named respectively after the resources they describe, present and describe each resource and their relation with the *lemmas* from P-AWL (in terms of recall).

#### 3.1.4.1 PAPEL

PAPEL[9] (Porto Editora's Associated Words – Linguateca) is a free lexical resource that focuses on word relations. It is a relatively recent resource, developed between September $1^{st}$, 2007 and December $31^{st}$, 2008. Currently, PAPEL is in version 2.0, available since March 2010. This resource is materialized by a text file representing in each line one relation with the format <word$_1$><relation><word$_2$>.

Table 3.5 provides information about the number of different words (*lemmas*) by POS and Table 3.6 provides information about the types and amount of relations that are available in the current version, with examples directly extracted from the resource (adapted to English).

Apart from the relations already discussed, in which some question types will rely on (synonym, hyponym/hypernym), PAPEL supports others that fit in the concept of meronymy/holonymy. Rela-

[9] http://www.linguateca.pt/PAPEL

| POS | Count |
|---|---|
| Adjective | 18,933 |
| Noun | 55,372 |
| Verb | 24,089 |
| Adverb | 1,389 |
| Total | 99,783 |

Table 3.5: Number of *lemmas* per part-of-speech for PAPEL.

| Relation | Count | Example |
|---|---|---|
| Synonyms | 79,035 | *Loyalty - Honesty* |
| Hypernyms | 61,477 | *Game - Fencing* |
| Origin Of | 816 | *America - American* |
| Part Of | 14,676 | *Wing - Plane* |
| Cause | 7,963 | *Fascinating - Fascination* |
| Producer | 1,278 | *Oak - Acorn* |
| End | 8,396 | *Hurt - Gun* |
| Manner | 1,245 | *Humbly - Humility* |
| Relative | 20,766 | *Tuned - Work well* |
| Total | 195,652 | |

Table 3.6: Relation types and examples for PAPEL.

tions of meronymy denote a constituent part of, or a member of something. Holonymy relations are the opposite of meronymy ones (for example, “has part” and “has member” relations). Synonym and hypernym relations can be extracted directly from this resource and hyponym relations can be formed inverting the elements in a hypernym relation.

#### 3.1.4.2 MWN.PT

MWN.PT (MultiwordNet of Portuguese) is a lexical resource shaped under the ontological model of *wordnets*. This resource focuses on relations as PAPEL does. It is available since May 2008, and was the first publicly available wordnet for Portuguese. It is developed and maintained by the NLX-Natural Language and Speech Group at the University of Lisbon, Department of Informatics. Apart from the content, the main difference between PAPEL and MWN.PT is the presentation method. MWN.PT is an ontological model representing relations in a hierarchical manner, in which words are grouped in synonym sets, called *synsets*, that establish a semantic relationship between them.

According to the project’s website[10], MWN.PT “spans over 17,200 manually validated concepts/synsets, linked under the semantic relations of hyponymy and hypernymy. These concepts are made of over 21,000 word senses/word forms and 16,000 lemmas from both European and American varieties of Portuguese. (...) It includes the subontologies under the concepts of Person, Organization, Event, Location, and Art works (...).”

[10] http://mwnpt.di.fc.ul.pt/features.html#main (last visited in July 2010)

As an example, an access to the MWN.PT with the word "carro" using VisuWords$^{TM}$ ("car" in English) generates the graph in Figure 3.1. "carro" is described as a word for "automóvel" ("automobile") and for "carro" ("car") itself. The full green triangle represents that "carro" is a kind of "veículo a motor" and "veículo motorizado" ("motor vehicle").

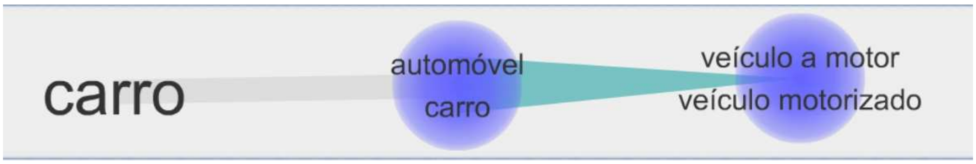


Figure 3.1: Graph generated by MWN.PT for the word "carro".

Table 3.7 presents the relations available in the MWN.PT resource, along with some examples.

| Relation | Count | Example |
|---|---:|---|
| Synonym | 4,309 | *Individual - Person* |
| Hyponym | 30,010 | *Person - Adult* |
| Hypernym | 17,932 | *Person - Living Being* |
| Has Part | 805 | *Poem - Stanza* |
| Is Part Of | 400 | *Cellule - Organism* |
| Has Substance | 35 | *Mineral - Rock* |
| Is Substance Of | 12 | *Vegetal - Vegetal Tissue* |
| Has Member | 5,929 | *Gangster - Gang* |
| Is Member Of | 11,220 | *Spouse - Couple* |
| Total | 70,652 | |

Table 3.7: Relation types and Examples for MWN.PT.

As PAPEL, this resource gives support to synonym, hyponym/hypernym, and meronymy/holonymy relations. Unfortunately, MWN.PT does not give support to any more different types of questions than PAPEL does. It does not have example sentences, definitions or antonyms. It should be also noted that MWN.PT covers only nouns. Despite this, its use widens the number of relations available, allowing for more diversity and content.

#### 3.1.4.3 TemaNet

TemaNet[11] is a free lexical resource that, like MWN.PT, is shaped under the ontological model of *wordnets*. TemaNet is being developed for Instituto Camões under the scientific coordination of Palmira Marrafa. It is a very recent resource but comes to this research as an essential one.

Like MWN.PT, TemaNet is arranged in a way as to group words in *synsets*. Apart from this already established semantic relationship between words, TemaNet divides all the *lemmas* in twelve semantic

[11] http://www.instituto-camoes.pt/temanet/inicio.html

domains. Table 3.8 presents the categories that form this resource and the number of concepts each one includes. Each of the semantic domains is a different *wordnet*.

| Semantic Domain | Count |
|---|---:|
| Art Works | 751 |
| Clothing | 712 |
| Communication | 737 |
| Education | 774 |
| Food | 1,711 |
| Geography | 1,487 |
| Health | 1,700 |
| Housing | 842 |
| Human Relations | 579 |
| Organisms | 2,193 |
| Sport | 698 |
| Transportation | 765 |
| Total | 12,949 |

Table 3.8: Number of words per semantic domain in TemaNet.

TemaNet represents a set of relations (synonymy, hyponymy/hypernymy, meronymy/holonymy), for each of the domains, widening the number of relations available. Additionally, this resource specifies definitions and example sentences, that could support the generation of definition and *cloze questions*.

In terms of the definition questions, this resource appears to be well suited for the task. It provides a *gloss* (a brief summary of a word's meaning) that allows the generation of definition questions. Since it is usually a short definition, one has to be careful while writing the instructions for those questions, allowing a certain degree of uncertainty of the terms (for example, using "What word has the closest meaning to the following definition").

Regarding *cloze* question automatic generation, unfortunately, this resource does not provide a good source of information. The example sentences lack in length and context establishment - a problem that has already been addressed in previous sections. For instance, the example sentence for the word "car" in TemaNet reads as follows: "She has changed her car three years ago." This does not provide enough information for the testee to answer correctly, since, instead of car, one could imagine an infinity of other objects.

### 3.1.5 Final Architecture

Conceptually, to support the proposed solutions two databases were used. The first one, a REAP.PT specific database, was used to maintain the state of the system during execution and use. The second one is responsible for storing the content of the lexical resources (relations, definitions, etc) and is context free, i.e., contains every possible data the resources provide about any word (looking forward for possible modifications of P-AWL or the inclusion of new features). A short version (with the relevant tables and fields) of the schemas of both databases is presented in Figures 3.2 and 3.3 for the REAP.PT database and resources database, respectively.

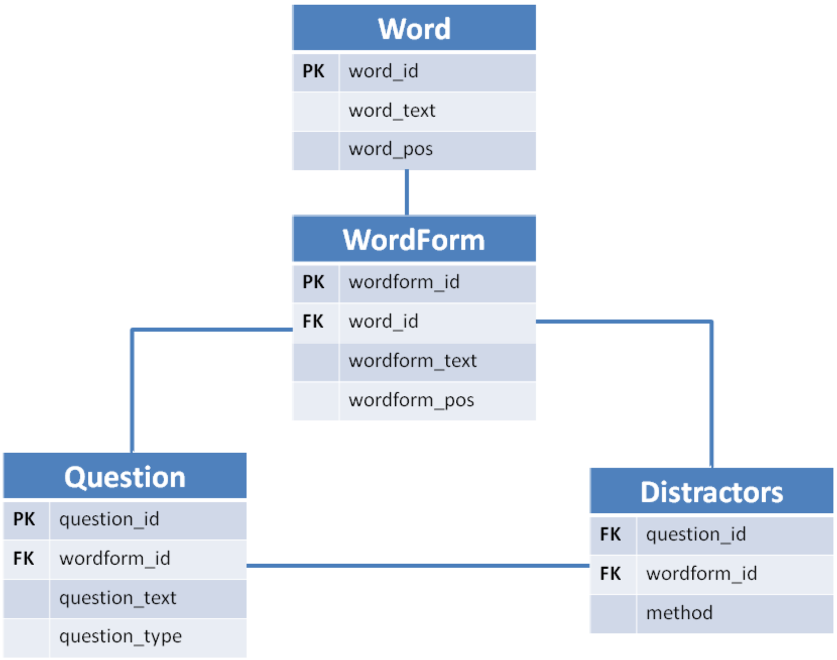


Figure 3.2: Detail of main database for REAP.PT schema.

One must pay attention to a few aspects regarding the databases. In the first place, the REAP.PT database is oriented to the concept of the *focus words*. The table "Word" contains the *lemmas* of the words, and the table "WordForm" contains the inflections for those same *lemmas* (described in Section 3.1.1). In the second place, the table "Question" is related with the table "WordForm", i.e., the questions are asked about the word inflections, which are related with their own *lemmas*. One should notice that all the words in the "Word" table are actually in the "WordForm" table as well.

Regarding question generation, it is impossible to generate proper *stems* and relations, taking into account all the requirements they should obey, in real time, i.e., generate proper information for the questions while the student is being assessed. This fact leads to the need to generate these artifacts previously and store them in the database, associated with the word they aim to test. This storing function is performed by the already mentioned "Question" table.

The "Distractors" table, as the name states stores sets of *distractors* for each question. The generation of these *distractors* will be described in the following sections.

It should be clear that, with this new setup, one is able to navigate between the main concepts of REAP.PT and easily present the correct content to the student (either documents or questions) based on the words that are being tested.

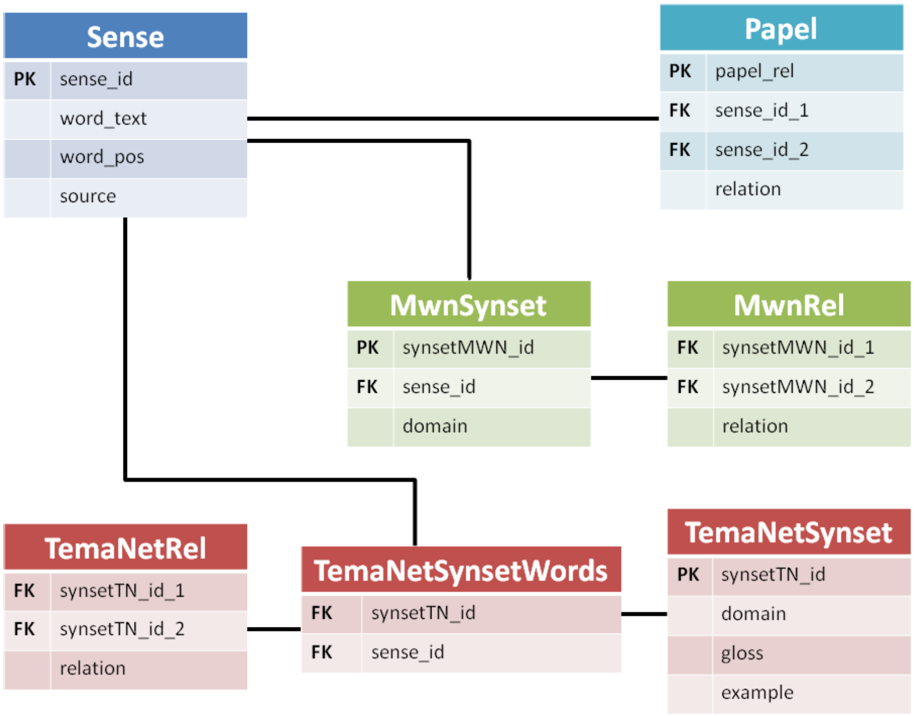


Figure 3.3: Resources database for REAP.PT schema.

Regarding the resources database, it now contains all the information available in the tree resources that were presented – PAPEL, MWN.PT and TemaNet. There is one common concept to all the resources represented in the table "Sense". This table has all the words that are present in the several resources. Notice that there can be several entries with the same word and POS since, for example, two resources can describe the same word. What one can not do is to represent these entries as having the same sense because there is no proof that they actually correspond to the same interpretation of the word. That is the reason why, in the table "Sense" there is a specific field to identify which one of the resources gave origin to that specific entry. Afterwards, it is possible to search on the "Sense" table for a specific pair word-POS and find which resources contain information for that pair.

Let us now focus on the particular representation of each resource:

- **PAPEL** is represented with a table with the same name. The content of PAPEL is stored as a simple relation between two senses from the "Sense" table;
- **MWN.PT** is slightly more complex since it is represented in its original form as a *wordnet*. The main concept here is the *synset*. The table "MwnSynset" is responsible for maintaining the relation between the senses (from the "Sense" table) and the MWN *synsets*. Having stored the *synsets* (and

as consequence, having represented synonym relations), the table “MwnRel” stores the relations that MWN.PT specifies between synsets;

- **TemaNet** has additional information – for each *synset* specifies a domain, a brief definition (*gloss*) and an example sentence. To replicate this data for every word of the *synset* is too expensive in terms of memory (in MWN.PT the only replicated info was a domain tag). So there is a table to group senses in *synsets* (“TemaNetSynsetWords”), and a table that adds new information to each *synset* (“TemaNetSynset”). Finally, like in the MWN.PT case, there is one table responsible for storing relations between *synsets* (“TemaNetRel”).

With all the resources in the database, it is now possible to relate the contents from the resources with P-AWL words. Figure 3.4 presents the recall of PAPEL, MWN.PT and TemaNet regarding P-AWL words, organized by POS, indicating the total in the end.

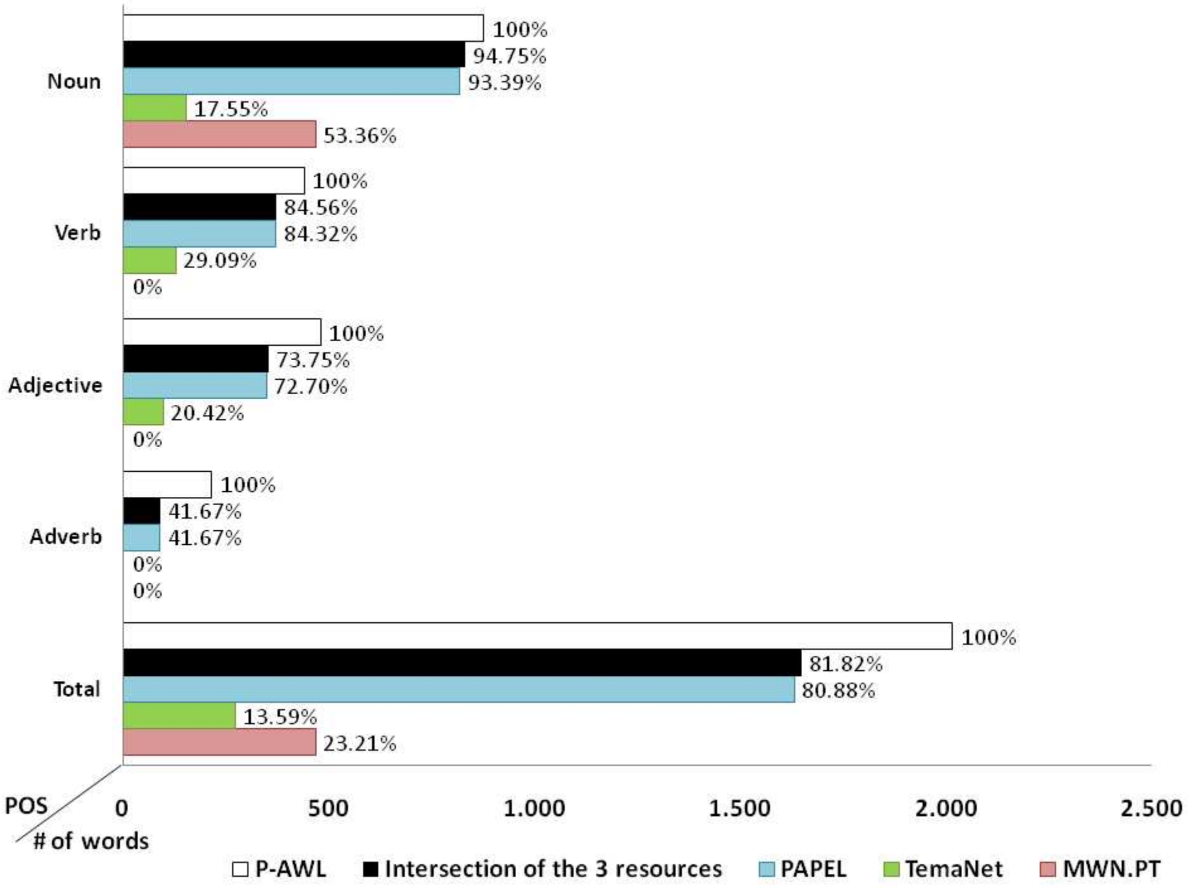


Figure 3.4: Recall of PAPEL, MWN.PT and TemaNet of P-AWL words, with POS discretization.

From the figure one can conclude that PAPEL is the only resource that provides adverbs, with a recall rate of 41.67% of P-AWL's adverbs. On the other hand, regarding verbs, PAPEL has a coverage of 84.32%. For this same POS, TemaNet expands 0.24% for a total coverage of 84.56%, i.e., 99.17% of the verbs in TemaNet are also present in PAPEL. For adjectives, the main contribution is again from PAPEL (72.70%) with a contribution of 1.05% from TemaNet (94.86% of the adjectives in TemaNet are also in PAPEL). For nouns and for the total contributions (with no POS discretization) let us attend on Figures 3.5 and 3.6. In each figure one can see which parts of P-AWL are covered by each resources or combination of resources, for nouns and for the total, respectively.

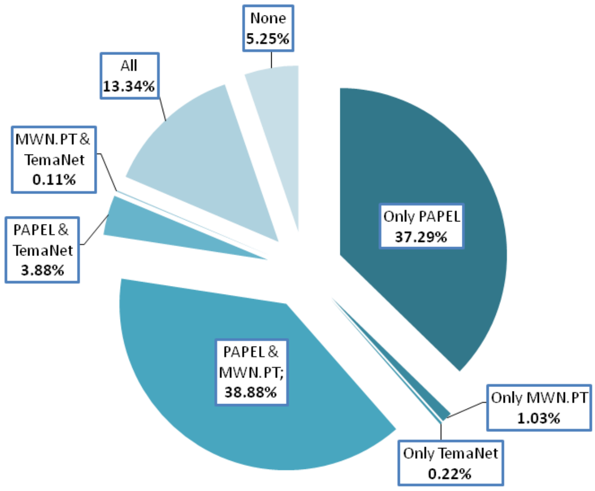


Figure 3.5: Contribution of each resource for the nouns in P-AWL.

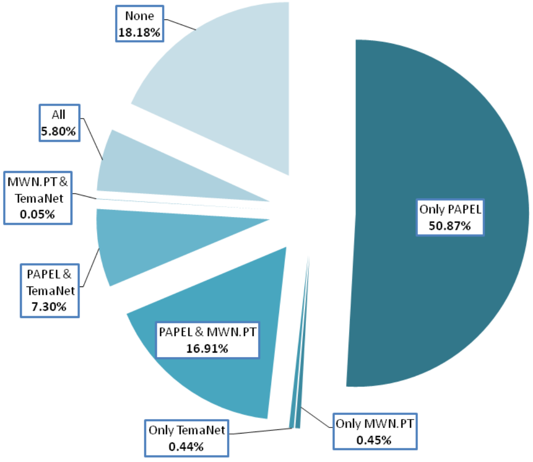


Figure 3.6: Total contribution of each resource for the total of P-AWL's words.

As one can see, PAPEL is an essential resource to support the question generation task. Alone, it provides relations for more than half (50.87%) of P-AWL's words. MWN.PT, on the other hand, proves to be a good resource regarding nouns. One should keep in mind that the intersections between resources do not represent repeated information, but complementary. In sum, there are still about 370 words (18.18%) that are not covered in any resource and, in view of this, the task of adding new resources that provide better coverage of P-AWL's words is still relevant.

## 3.2 Question Generation

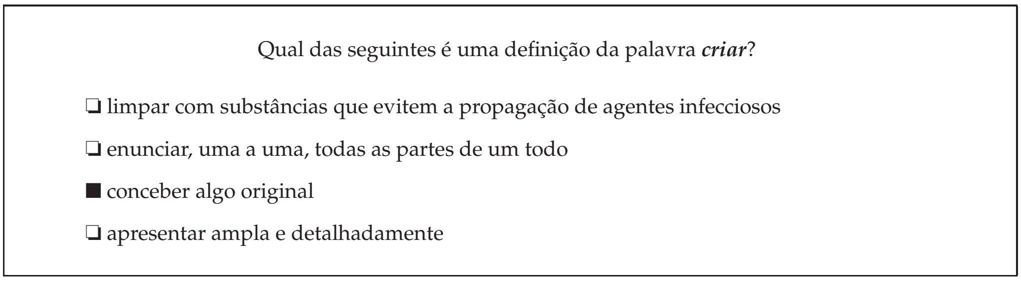
With the new information, REAP.PT has the necessary information to generate questions automatically. The following sections focus on specific types of questions, describing how they were developed and presenting examples.

### 3.2.1 Definition Questions

Since the only resource that provide definition texts, TemaNet, has such a low recall of P-AWL words (13.59% – see Figure 3.4), one tried to extract the definition texts from the same source where the dictionary lookups collect the information to present to the student – the Porto Editora's online dictionary. This dictionary has more than 920,000 words, thus being an excellent resource to extract definition texts. If a particular word does not exist in this dictionary, the system then tries to extract a definition from TemaNet.

Usually a dictionary entry has more than one definition for the same word, each one representing a sense of the word. At the same time, these definitions tend to be ordered from the most to the least common sense in the language. For that reason, and since there is no mechanism for disambiguation yet, the system chooses to use the first definition that is available. For this type of questions, *distractors* are generated randomly using the definitions of the remaining P-AWL words.

Figure 3.7 present a real definition question generated automatically by the REAP.PT system for the word "rota" (in English, "route").



Qual das seguintes é uma definição da palavra ***criar***?

❏ limpar com substâncias que evitem a propagação de agentes infecciosos

❏ enunciar, uma a uma, todas as partes de um todo

■ conceber algo original

❏ apresentar ampla e detalhadamente

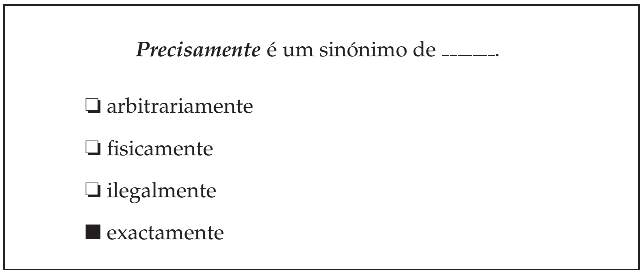
Figure 3.7: Real example of an automatically generated definition question.

### 3.2.2 Synonym and Hypernym/Hyponym Questions

Synonym, hypernym and hyponym questions are very similar when it comes to generation, and for that reason are described in the same section.

To generate questions of these types, REAP.PT first searches through all the resources and extracts all the occurrences of a particular relation (depending on the type of question) for the *focus word* being tested. After this step, it tries to see if there is the same relation, between exactly the same words, in more than one resource. This duplication may indicate that a particular relation is the most common for a given word, fact that leads the system towards the usage of this particular relation to generate the question. If no such duplication exists, all there is to do is to choose a random entry and use it to compose the question. Regarding *distractor* selection REAP.PT selects only words with the same POS of the *focus word*, and also tries to choose *distractors* that have the same level classification of the word being tested.

Figure 3.8 presents a real example generated by REAP.PT of a synonym question.



***Precisamente*** é um sinónimo de _______.

□ arbitrariamente

□ fisicamente

□ ilegalmente

■ exactamente

Figure 3.8: Real example of an automatically generated synonym question.

### 3.2.3 Cloze Questions Distractors

As already mentioned, one will use the *stems* that were developed at University of Algarve to serve as basis for the generation of *cloze* questions. The main focus was on the generation of a coherent set of *distractors*.

For that purpose, a preliminary study was carried out in order to understand the testees' behavior when presented to open *cloze* questions. After the description of this preliminary study the present document will focus on the experimental setup that was developed and the evaluation of the experience.

#### 3.2.3.1 A preliminary experiment

A preliminary experiment was carried out among 4 test subjects, using a set of 100 randomly selected sentences, distinct from the set used for the main experiment. They were asked to complete each sentence, not being provided any answer choices. With this experiment, one expected a low percentage of correct answers (answering exactly the target word that a specific stem aims to test), and were interested in finding the most common reasons for "incorrect" answers.

The subjects did not participate in the test in exactly the same conditions:

- subject A is a native speaker and had no previous knowledge of the sentences;
- subject B does not consider her/himself as native-speaker, and had only partial knowledge of the corpus of sentences, since he/she has been involved in their selection;

- subject C and D are Portuguese native speakers and they also had only partial knowledge of the corpus of sentences.

Since team members B, C and D were involved in the selection of sentences, a random set was retrieved from the sentences selected by each member, so that only 25% of the sentences were in fact previously known to each person. For each response, there were three possible outcomes: unable to find a coherent word to complete the stem, able to find exactly the expected word and able to find a word but not the expected target word. Results are shown in Table 3.9:

| | A | B | C | D |
|---|---|---|---|---|
| No Response | 4 | 23 | 23 | 24 |
| Correct | 2 | 14 | 27 | 20 |
| Incorrect | 94 | 63 | 50 | 56 |

Table 3.9: Distribution of the answers in the preliminary experiment.

While subjects B, C and D, with previous, even if partial, knowledge of the corpus of sentences were able to provide correct answers for 14, 27 and 20 sentences, respectively; subject A only got 2 answers right. This seems to confirm that previous knowledge of the sentences, even if diluted among the large number of each subject's selected sentences, and the time gap between the selection and the testing (over a few months), may influence their response.

The attitude towards the test has also been different among the participants. Subject A only left unanswered 4 questions, while B, C and D were much less assertive or were in fact unable to find adequate answers for 23 or 24 questions.

The analysis of the "incorrect" answers is illustrative of the cognitive mechanisms involved in the test and may shed some light on both the quality of the sentences and the task difficulty. It should be kept in mind that having chosen a word that is not the target word does not mean that it is inadequate.

This study revealed some problems associated with the use of open *cloze* questions (when no set of answer choices is provided). While it would be impossible to go through all cases here, one will highlight some remarks:

- incorrect answers may arise from the choice of synonyms or antonyms (notice the case where the target word is neutral (for example "vary"), and incorrect answers correspond to the positive ("increase") and negative ("reduce") polarity);
- hyponymy and hyperonymy relations often compete with synonyms/antonyms as incorrect answers;
- in some cases, a specific full verb is replaced by a support (or "light") verb, practically devoid of meaning (for example, "give" instead of "confer");

- less frequent adverbial quantifiers (like "marginally") are replaced for equivalent adverbs but with a broader selection ("very" and "little");
- collocation patterns arise in the choice of the answers that do not match the target word.

The use of multiple-choice questions with automatic *distractor*'s generation is able to solve some of the aforementioned problems, such as the use of synonyms, antonyms, hyponyms and hyperonyms as answers (using, for instance, lexical resources to eliminate these choices).

#### 3.2.3.2 Experimental Setup

For this new experiment, 20 stems were randomly selected and, for each, six sets of *distractors* were generated (each element of a given set was generated with the same generation strategy as the remaining ones), thus yielding 120 sets of *distractors*. This setup allows one to isolate each generation method thus being capable of draw conclusions over each method, separately. Each test subject was asked to answer 10 *cloze* questions. The questions were randomly chosen, maintaining a balance over the pairs *stem-distractors* that were already answered, and avoiding the use of repeated stems in each test.

The test subjects, also divided by native and non-native speakers, were asked to select all the words that could fit in the stem from the set of words provided. One had 247 participants in our test (212 native and 35 non-native speakers).

The present study focus on 4 main generation methods of *distractors*: manual, random, graphemic, and phonetic *distractors*. The remaining 2 sets of distractors are generated with the random and graphemic methods but applying lexical resources in order to filter out synonyms, hyponyms and hyperonyms of the target word, which, if included, might also fit the stem as correct answers and fail to function as *distractors*. One will now focus on each method in particular.

For a set of 20 randomly selected stems, a set of *distractors* was **manually** produced by the team members. These *distractors* were selected based on the following criteria:

- quasi-synonymous or quasi-antonymous words, which do not correspond exactly to the negative/positive overall sense of the target word in the sentence;
- similar spelling or similar sounding words;
- false-friends, taking as competing languages the pair English/Portuguese;
- (pseudo-)prefix and suffix variation.

For example, for the target word *condução*, in the sense of driving a vehicle, the manually selected *distractors* were:

- *direcção*: noun derived from the verb *dirigir* (drive) that is a perfect synonym of *conduzir* from which the target word is derived (notice, in this case, that even if the two verbs are synonyms, their derived nouns are not, and cannot fit in the stem);
- *condição*: phonetic/graphemic similarity; and
- *redução*: pseudo-prefix re-/con- variation, but otherwise semantically unrelated words.

The second approach was to generate *distractors* in a **random** way. This method was developed to settle a baseline on automatic *distractor* generation. Despite the name, this method is not completely random. The *distractors* were chosen using only the P-AWL entries with the same POS and level.

The third developed set of *distractors* was the set of **graphemic** *distractors*. To compute this set of *distractors*, the P-AWL words with the same POS were used, and the Levenshtein Distance [25] was computed from the target word to each word, using unit costs for the three operations (deletion, insertion and reversal). The words with the lowest distances were selected, among the ones which provided a distance lower than five. The set of *distractors* thus obtained was completed by recurring to words with a lower distance from a different POS (ending up with gender or number variations for nouns, or tense and person for verbs).

**Phonetic** *distractors* were the fourth approach. This method of generation tries to explore the most common spelling errors for the Portuguese language. To accomplish this task, a table of common mistakes was used. For example, "ss" is frequently confused with "ç" (before "a", "o" and "u") and "c" (before "e" and "i"); "j" can be in some cases confused with "g" (before "e" and "i"); etc. The target word is submitted to this table of modifications and several (misspelled) words are thus obtained.

The *leia* grapheme-to-phone tool [9] was used to provide phonetic transcriptions, so that misspelled words sharing the same transcription as the target word might be selected.

For example, the Portuguese word *começar* (to start), shares the phonetic transcription (/kum@s"ar/, using SAMPA symbols) with the misspelled words *cumeçar*, *comessar*, etc.

Finally, lexical resources may improve methods of automatic distractor generation by **filtering** out correct candidates choices, other than the target word. One used two different resources two support this task: PAPEL and MWN.PT. These two resources were used to generate the *Random + Filtering* and the *Graphemic + Filtering* categories of *distractors*.

- PAPEL – PAPEL supports synonym, hyponym and hyperonym relations and has a recall of 80.93% of the P-AWL vocabulary. PAPEL was used to exclude *distractors* with this type of relationship with the target word. For instance, for the target word *refinação* (refinement), our graphemic method produced the distractor *realização* (implementation) which was later discarded with the aid of PAPEL. Unfortunately, there is no support for antonym relations with this resource.

- MWN.PT – This resource also provides synonymy, hyponymy and hyperonymy relations, covering only nouns (recall of 53.36% of the nouns in P-AWL). However, MWN.PT catalogues each word in a specific domain. One used this property to exclude words within the same domain (for example, if two words represent an occupation – "lawyer" and "doctor" – they can easily fit in the same stem, if the context does not restrict the choice).

#### 3.2.3.3 Evaluation

Results obtained with different methods were compared with the results achieved with manually produced *distractors*. The latter may be seen as a sort of goal/best result that those methods would try to emulate. Table 3.10 presents the distribution of correct answers, i.e., answers in which only the expected choice was selected, across the type of *distractors*, for both Native (NS) and Non-Native (NNS) Speakers.

| Method | NS (%) | NNS (%) |
|---|---|---|
| Manual | 87.9 | 70.0 |
| Random | 83.1 | 78.0 |
| Random + Filtering | 84.3 | 73.8 |
| Graphemic | 83.2 | 61.0 |
| Graphemic + Filtering | 87.0 | 86.0 |
| Phonetic | 88.3 | 75.7 |
| Total | 85.6 | 74.2 |

Table 3.10: Percentage of correct answers in the *distractors* experiment.

As one would expect, native speakers tend to achieve a higher percentage of correct answers than non-native speakers. This result is consistent for all our *distractor* generation methods.

Variation due to the different methods does not seem to affect the results of native speakers in a significant way. The standard deviation with native speakers is 2.37 while with non-native speakers is 8.37, that is, about three times higher. This seems to indicate that non-native speakers are more prone to produce incorrect answers, depending on the *distractor* generation method than native speakers, who can activate their language knowledge to detect other clues in the stem in order to find the best match.

Concerning native speakers, the fact that the results for the phonetic method are slightly superior (88.3%) can be easily justified since they are aware of the spelling rules for their language. In fact, having sets of *distractors* composed only by this method makes it easier for native speakers, since they become really focused on the task of finding the misspelled words. As it would be expected, the lowest result was obtained with the random *distractors* (83.1%), and even so it is only 4.8% (absolute) less than the manual method.

Results for non-native speakers are on average 11% lower than those for native speakers. The highest difference occurs for the manual *distractors* (17.9%). This difference may be taken as a confirmation of the adequate selection of the manual *distractors*. The random and phonetic methods show similar results (78.0 and 75.7%, respectively). However, when both are compared with the performance of native speakers, the phonetic method reveals its interest for *distractors* generation, since it provides significantly less correct answers (less 5.1 and 12.6%, respectively). The graphemic method exhibits the lowest number of correct answers for non-native speakers (22.2% less than the results for native speakers). This result indicates that the confusion introduced by presenting similar (valid) words may be the most important cause for the incorrect answers, and that this method might be particularly suited for screening different levels of progression in vocabulary acquisition.

For native speakers, the filtering version of the generation methods always produced higher correction rates (1.2% more than the random method and 3.8% than the graphemic method). This positive, if small, impact in performance may also means that it became easier (even if only a little easier) to rule out incorrect answers.

It is also important to focus on the number of choices testees selected. Non-native speakers always selected only one choice, a fact that can be explained by their concern to answer correctly question right, stopping when they have found a correct answer (in their notion). For native speakers we registered a very low percentage of selection of more than one word, across the several types of *distractors* (0.29% for the random, random filtered with resources and phonetics categories, 0.34% for graphemic filtered with resources and 0.57% for the graphemic ones). The manual category, as expected, did not generate any confusion.

Several native testees commented on the level of the stems themselves. Although they were randomly selected among a set of manually revised sentences, care was not taken to ensure that the level of the sentence with a blank space fitted the level of the target word. In fact, a later analysis revealed that 70% of the sentences had a higher level than the corresponding target words (on average, 2.4 higher than the target word level, with a standard deviation of 1.6). This was a problem for non-native testees, who in some cases had to look up unknown words in a dictionary.

# 4 Extending the General Functionality

The tasks described in the previous chapter (Chapter 3) endowed REAP.PT with consistently connected information (*focus words*, inflections of the *focus Words*, questions, etc.) This fact allowed one to extend the general functionality of the system, i.e., redo some tasks maintaining integration between elements (Section 4.1), and adding new features to the system (Section 4.2 and Section 4.3).

## 4.1 New Set of Documents

One of the major problems with the REAP.PT system was the lack of a set of documents to be presented to the student that met the goals of the system itself, i.e., a set of recent documents, correctly filtered and classified, containing words from the P-AWL.

The corpus that was beeing used previously was the WPT05 Web corpus [12]: a collection of over ten million documents obtained by the crawler of the Tumba! search engine (called Viúva Negra [26]), produced by the XLDB Node of Linguateca. As the name states, this corpus was collected in 2005, and for that reason does not suit REAP.PT anymore, since the documents are outdated. On the other hand, the formats in which the WPT05 collection is available (RDF/XML and ARC) turned out to be difficult to process and to present documents in an appealing way for the student.

Thus, the integration of a new set of documents in the system was one of the main tasks to accomplish during this thesis. In the present section, one presents the new set of documents (ClueWeb09), a description of the filtering process and the results of the task.

### 4.1.1 ClueWeb09

ClueWeb09 is a collection of 1 billion web pages, in ten languages, that was collected between January and February 2009 by the Language Technologies Institute at Carnegie Mellon University. As the web page[13] of this resource describes, it was created "to support research on information retrieval and related human language technologies". The total size of the dataset occupies about 6 terabyte (TB) compressed.

[12] http://xldb.fc.ul.pt/wiki/WPT_05_in_English (visited in Jul. 2010)

[13] http://boston.lti.cs.cmu.edu/Data/clueweb09/ (visited in Jul. 2010)

The format in which the Web pages are stored is the WARC file format [27]. This format evolved from the ARC format, containing the data represented as HTML requests (with metadata). In the ClueWeb09 each WARC file is about 1 gigabyte (GB), uncompressed, containing several tens of thousands of web pages.

ClueWeb09 is also organized according to a language identification module. Table 4.1 provides information about the languages and corresponding number of pages in the dataset.

| Language | Number of Pages |
|---|---:|
| English | 503,903,810 |
| Chinese | 177,489,357 |
| Spanish | 79,333,950 |
| Japanese | 67,337,717 |
| French | 50,883,172 |
| German | 49,814,309 |
| **Portuguese** | **37,578,858** |
| Arabic | 29,192,662 |
| Italian | 27,250,729 |
| Korean | 18,075,141 |

Table 4.1: Languages distribution in the ClueWeb09 corpus.

Therefore, in the REAP.PT context one used a subset of the ClueWeb09, comprising only the portion of the Portuguese documents (about 160 GB compressed).

### 4.1.2 Adapting the Chain of Filters

To process the new corpus one needed to do some modifications to the chain of filters already developed in the first porting of the REAP system that was briefly described in Chapter 2. There were changes made in the order of execution of the filters, in the filters themselves, in the data representation, and in the methods used to process the whole corpus.

One of the problems of the filter chain was the bias caused by the fact that the whole page was being processed in the same way, i.e., there was not a content breakdown regarding the several types of elements that a web page is composed of (menus, comments on blogs, captions, etc.). This turned out to be a problem, particularly when feeding the classifiers (readability and topic) with the complete document. For instance, if a web page has a navigation menu (such as music, sports, science, etc.) the topic classifier would mistankenly consider these words to classify the page.

Therefore, one needed a mechanism to extract the main content of a web page, discarding the entire side information. For this purpose, we used a Java library called Boilerpipe written by Christian Kohlschütter [28] and released under the Apache License 2.0. The official web page of the project[14]

[14] http://code.google.com/p/boilerpipe/ (visited in Jul. 2010)

describes this library as a set of “algorithms to detect and remove the surplus “clutter” (boilerplate, templates) around the main textual content of a web page.”. Apart from this feature (extracting the main content), Boilerpipe also provides an extraction mechanism that simply clean all the HTML tags and remaining control information (such as JavaScript), leaving only the textual information that the page contains.

These two features allowed a new representation of the documents. Each document is now represented in two ways: a version with the complete text information in the page, and a second version with the main text content only. Each filter and classifier can now take advantage of this fact, using the appropriate version of the document for its processing task.

The second major change in the filter chain has to do with the return value of each filter. The past version of the chain allowed only for the filter to return a boolean value: either the document passed the filter or not. This is sufficient for the documents that do not pass the filter but, for the remaining ones, there is several information computed by the filter, that is being lost and is useful in the context of the REAP.PT system.

Having this new set up in mind, let us now focus on each filter of the chain, and how it takes advantages of the new features. Filters will be presented in the order they are executed in the new setup of the chain.

1. ***Less Than Minimum Number of Words* Filter** – this filter is intended to remove small documents (in the REAP context, documents with less than 300 words). It uses the version of the document with the main text only since one wants to estimate the number of the words the student should read. As it counts the words of the document to decide if it passes the filter or not, if it does, the filter also returns the number of words of the main text in the document. This information is stored in the REAP.PT database being used by some features of the system (as the teacher’s search interface – see Section 4.2);

2. ***Profanity Words* Filter** – this filter removes documents that contain obscene language that is defined in a list based on the *Open dictionary of slang and idiomatic expressions*[15]. Given the goal of this particular filter it uses the complete version of the document, in order to detect profanity words in the entire document;

3. ***Document Containing Just List of Words* Filter** – this filter removes documents that do not have a significant text section, containing only a list of words. To do so, one compares the proximity of each document with a reference document (an extract of CETEMPublico[16]). Representing each document as a vector of POS 3-grams for the key, and the number of occurrences of the 3-gram

[15] http://natura.di.uminho.pt/ jj/pln/calao/calao.dic (visited in Jul. 2010)

[16] http://www.linguateca.pt/cetempublico/

as the value, it computes the documents' proximity using Cosine Similarity (Equation 4.1). Since it aims at analysing the format of the document, this filter is fed with the entire version of the document;

4. ***Lack of Focus Words* Filter** – this filter removes documents that do not have, at least three of the *focus words* that P-AWL defines. The order in which this filter appears in the chain was changed for two main reasons. In the first place, given the expansion of the *focus words* set with the inflections of the original *lemmas* of the P-AWL there are almost no documents that do not pass this filter (for instance, one is considering all the verb tenses and gender/number variations). In the second place, this filter, receiving as input only the main content of the page, is now being used to identify which words (from the extended *focus words* set) are in which documents. To accomplish this it has to process the entire text (not stopping if it finds three words from the list). This is obviously a task that consumes a great amount of time, and for that reason should be carried in documents that will be added to the accepted documents set. This new process covers the problem of identifying the *focus words* in the documents – essencial to the workings of the REAP.PT system;

5. ***Readability* and *Topic* Classifier** – both classifiers are now fed with the main content of the document provided by Boilerpipe, eliminating the bias problem caused by menus and other characteristic web page elements.

$$cos(d^a, d^b) = \frac{\sum_{i=1}^{n} d_i^a \times d_i^b}{\sum_{i=1}^{n}(d_i^a)^2 \times \sum_{i=1}^{n}(d_i^b)^2} \tag{4.1}$$

### 4.1.3 Processing the ClueWeb09

The WPT05 corpus had been processed using the Hadoop[17] MapReduce [29] framework. The main advantage of this particular framework over other High Throughput Computing (HTC) approaches is that it moves the computation to the data, which is faster than moving the data itself. It turns out that there were reading problems when trying to process the entire corpus using Hadoop.

Since L$^2$F also maintains another HTC system, called Condor[18], we changed the approach of the processing task, modifying the front-end of the chain of filters, adapting it to run on this specific framework. Like Hadoop, "Condor is a specialized workload management system for compute-intensive jobs" providing "a job queuing mechanism, scheduling policy, priority scheme, resource monitoring, and resource management.".

There were **36,897,891** documents processed. 0.0076% of these documents were found to be badly formatted, and due to the insignificant quantity, were ignored. In the end, the REAP.PT system ended

[17] http://hadoop.apache.org/ (visited in Jul. 2010)

[18] http://www.cs.wisc.edu/condor/ (visited in Jul. 2010)

up with **1,827,299** documents. Figure 4.1 presents the percentage of the original set of documents that passed in each filter. Table 4.2 and Table 4.3 show the distribution of the documents according to their readability level and topic classification, respectively.

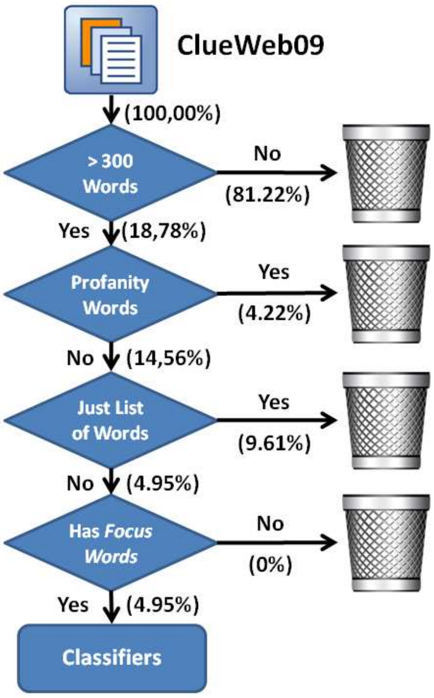


Figure 4.1: Filter chain schema and percentages of deleted documents in each step.

| Readability Level | Number of Pages |
|---|---:|
| Five | 7,461 |
| Six | 31,010 |
| Seven | 131,750 |
| Eight | 268,434 |
| Nine | 478,678 |
| Ten | 504,690 |
| Eleven | 285,594 |
| Twelve | 119,682 |

Table 4.2: Readability level distribution of the accepted documents.

### 4.1.4 Storage and Presentation

One of the main discussions that took place while developing this work was related with the methods used for storing and presenting the documents.

In what concerns storage, the biggest issue was the amount of space needed to store all the documents that passed the filters, taking into account the demand for fast access to each document. So far,

| Topic | Number of Pages |
|---|---|
| Economy | 6,840 |
| Education | 1,661 |
| Environment | 1,620 |
| Health | 191,016 |
| Justice | 857 |
| Meteorology | 0 |
| Politics | 7,290 |
| Security | 70,042 |
| Society | 13,899 |
| Sports | 40,217 |
| National | 181,429 |
| International | 99,801 |
| None | 1,212,627 |

Table 4.3: Topic distribution of the accepted documents.

the documents were stored in the REAP.PT database itself. This solution turned out to be unfeasible given the size of the processed corpus. To solve this problem, the processed set of documents was saved in the AFS (Andrew File System), granting reading permissions for the REAP.PT web page. In addition to the HTML source code, one stored also the main content of each document, for further reference. To successfully use these documents in the REAP.PT system one needs several metadata information for each document, such as readability level, topics, *focus words*, etc. This information was then stored in the database, allowing the system to have quick access to the main control criteria, when assigning reading tasks to a particular student. Figure 4.2 presents the new tables that were added to the database in order to accomplish this task.

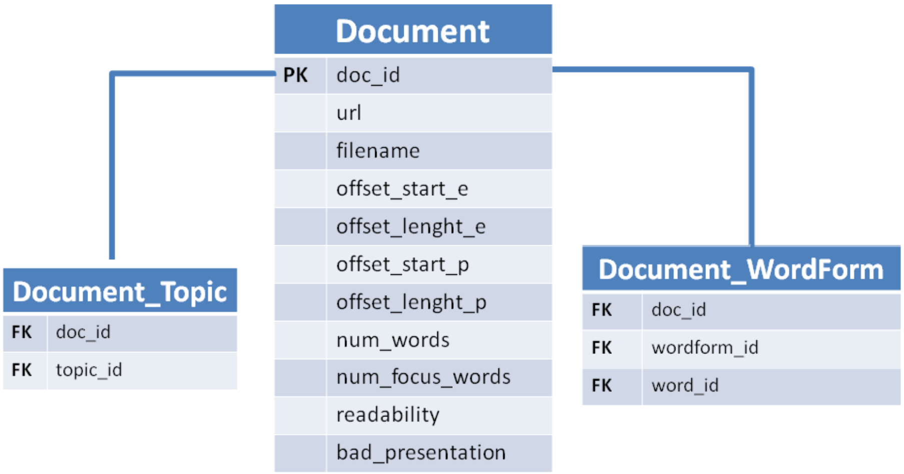


Figure 4.2: Tables in the REAP.PT database used to store the documents metadata.

There are three tables in the database that are used to store the documents metadata:

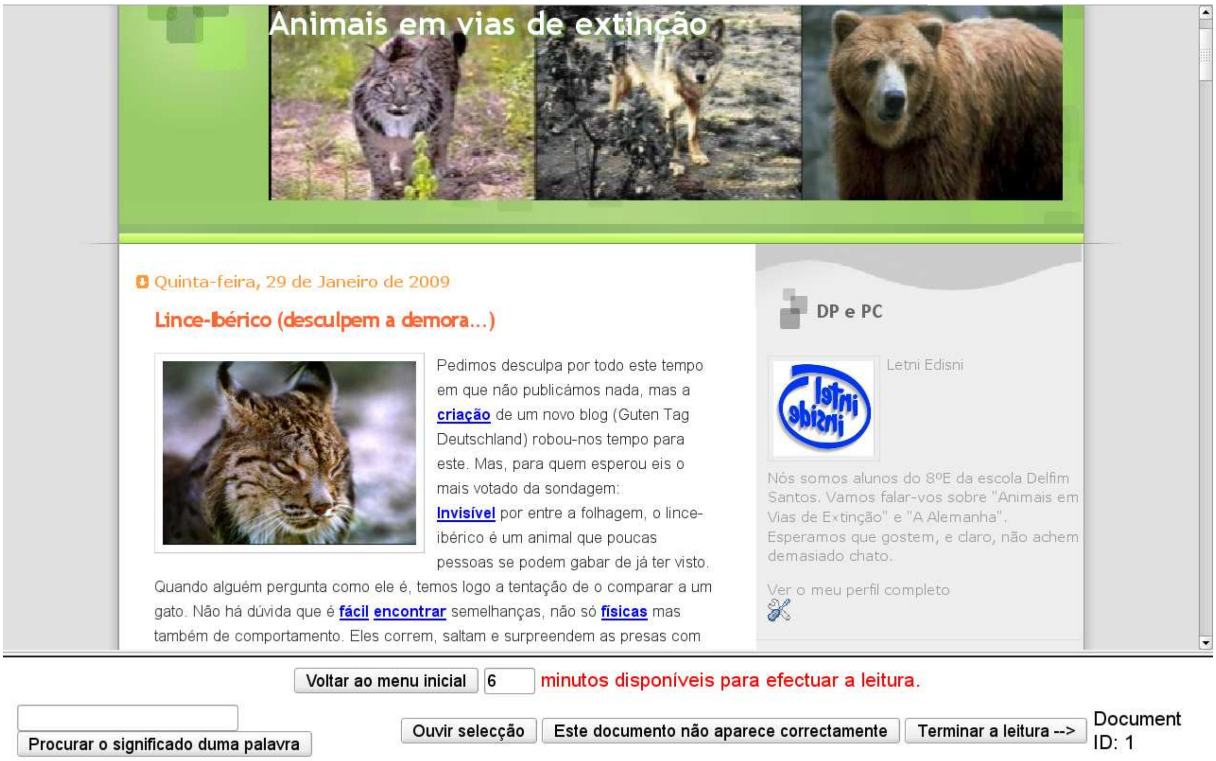
- ***Document_Topic*** – stores the information about the topics that were assigned to the document;
- ***Document_WordForm*** – stores the information about the *focus words* that are present in the document. This table has a reference to the *WordForm* table and another to the *Word* table, to have references of the words as they appear in the document but also about the *lemmas* of those words. This is important since, for example, if one wants to teach the word "dog", one may want to consider documents in which the word "dogs" appear;
- ***Document*** – stores several information about each document: its URL, the filename (location in the AFS file system), both offsets within the file (for the complete version of the document, and for the partial version) and corresponding length of each element, the number of words computed for the main part of the document, the number of *focus words*, the readability level, and finally a flag to mark the document as not being presented correctly.

This last attribute of the *Document* table leads to the presentation issue. As already said, this set of documents was collected in early 2009. This fact implies that many references that are in the HTML documents are not online anymore (such as images, CSS, etc.). Since one wants to remove the links on each page and highlight the *focus words*, forwarding the user to the actual web page (via URL) is out of question. Altogether, these facts may cause a bad presentation of the documents. For that reason, the user has the option to mark a page as badly presented, being redirected again to the page in the REAP.PT system where he/she can select another document. Later on, if a web page has been marked several times as being badly presented, the system can simply discard this page, not showing it again to the students. But not only students can classify the documents. In fact, the system allows the teacher to classify each document, seriously taking into account the classifications given by each teacher. In the section 4.2, one will explain the functionalities available to the teacher regarding document classification.

Figure 4.3 presents the interface of a well-behaved web page (where images and the style of the page were preserved). The control frame, where the user can lookup words in the dictionary, end the reading and listen to a part of the document, has now the aforementioned button to mark the document as badly presented.

Before actually seeing a document, the user has to choose which document he/she wants to read. Given the characteristics of the ClueWeb corpus, one added a new set of options that will be considered while showing a list of documents for the student to read. When defining the topics in which the student is interested, he/she has now two new options:

- **Domain Restriction** – ClueWeb does not make distinction between the varieties of Portuguese in the documents. For that reason, the student now has the option to be presented only with documents from the *.PT* domain, only *.BR* domain or any domain.



Animais em vias de extinção

Quinta-feira, 29 de Janeiro de 2009

Lince-Ibérico (desculpem a demora...)

Pedimos desculpa por todo este tempo em que não publicámos nada, mas a criação de um novo blog (Guten Tag Deutschland) robou-nos tempo para este. Mas, para quem esperou eis o mais votado da sondagem: Invisível por entre a folhagem, o lince-ibérico é um animal que poucas pessoas se podem gabar de já ter visto. Quando alguém pergunta como ele é, temos logo a tentação de o comparar a um gato. Não há dúvida que é fácil encontrar semelhanças, não só físicas mas também de comportamento. Eles correm, saltam e surpreendem as presas com

DP e PC

Letni Edisni

Nós somos alunos do 8ºE da escola Delfim Santos. Vamos falar-vos sobre "Animais em Vias de Extinção" e "A Alemanha". Esperamos que gostem, e claro, não achem demasiado chato.

Ver o meu perfil completo

6 minutos disponíveis para efectuar a leitura.

Document ID: 1

Voltar ao menu inicial
Procurar o significado duma palavra
Ouvir selecção
Este documento não aparece correctamente
Terminar a leitura -->

Figure 4.3: Reading interface.

- **Type Restriction** – many documents extracted from the ClueWeb have Blog content. It is a fact that Blog content is extremely sparse, and the quality of the text (syntactically and semantically) can be very poor. For this reason, the student can now also choose not to learn from Blog material. This is accomplished by simply discarding documents that have in their URL the term “blog”.

Additionally, the list of documents from where the student can choose his/hers next reading has now a brief extract of the document. With this information the student can make a more informed decision about the next reading that will be assigned to him/her.

## 4.2 Teacher Interface

Before moving to the field trials, it is crucial to give the teachers control over REAP.PT. For that reason, a *Teacher Interface* was created (adapted from the Carnegie Mellon University version).

This section focuses on two new features: the *Search Interface* and the *Document Viewing System*. Apart from these two, the teacher can add a new student to the system, reset his/hers password, and change his/hers own password.

### 4.2.1 Search Interface

The *Search Interface* allows the teacher to search for documents using four variables to control the search:

- **Readability level** – the user can establish a minimum and a maximum value for this parameter. The documents retrieved will have a readability level according to the defined boundaries. If the user does not fill this parameter, the system will search for documents in all reading levels;
- **Number of words** – establishing a minimum and a maximum for this parameter, the user will be presented with documents that contain a number of words that correspond to the indicated values. Again, if left in blank, the system will retrieve documents with any size;
- **Topic** – the default value for this parameter is ***ANY***, where the system will retrieve documents of any topic. The user has the option to specify a particular topic to search for;
- ***Focus words*** **–** the user is allowed to define a set of words, of any size, that he/she decides. These words must be contained in the set of words that REAP.PT recognizes as *focus words*. If there is any word that does not belongs to this set, it will simply be ignored. One should notice that the teacher can search for the *lemmas* of the words and for the inflections of those lemmas. The sistem will then retrieve any document that contain, at least, one of those words.

Of course, all of the parameters just described can be combined to build a search query to the database that will retrieve documents that satisfy all the values established. The result of the search is a list of documents, with the identifier of the document, its URL and a “preview” link that allows the teacher to see the document as it is presented to the students. If the user has selected some *focus words*, those will be highlighted in the “preview” version of the document.

This kind of feature can be extremely useful, for instance, to pre-select a specific document for group reading. The teacher can control what the class is reading and assure a good quality of the document and a correct sense of the word that is supposed to be taught.

### 4.2.2 Document Viewing System

Another interesting and crucial feature is the *Document Viewing System*. This system works as a solution to the presentation and filtering issues pointed out in Section 4.1. As has already been said, for a system such as REAP.PT to be accepted it is very important to give control to the teachers.

The *Document Viewing System* allows the teacher to navigate in the documents database that appears in this subsystem as a list. Each entry of the list, i.e., each document, is here represented with seven attributes:

- URL;
- Document identifier;
- Readability level;
- List of topics;
- Number of words in the document;
- Number of *focus words* in the document;
- List of *focus words*.

The user is also allowed to sort the list using some of the attributes presented, which are the identifier of the document, the readability level, the number of words and the number of *focus words* in the document, either in an ascending or descending way (clicking on the name of the parameter will sort the list in an ascending way according to that parameter and clicking again will cause the documents to be sorted in a descending way, for that attribute).

In the same list, associated to each document are two links: a "preview" link (with the same role as the "preview" button from the preceding section, but this time, highlighting all the *focus words* in the document) and "Rate/Exclude" link. The latter drives the teacher to a new window with two frames: a frame with a preview of the document (with all the *focus words* highlighted) and a bottom frame with a form. This form is the important feature of this *Document Viewing System*. The teacher has now the chance to manually rate the document that he/she is seeing and/or mark it for exclusion. The inputs of the form are:

- Quality of the document – a scale from 0 (very poor quality) to 10 (excellent quality);
- Readability level – a scale from 5 to 12 corresponding to the $5^{th}$ and the $12^{th}$ levels;
- Exclusion checkbox – where the teacher can mark the document to be excluded;
- Reason of exclusion (optional) – where the teacher can select the reason why he/she has marked the document for exclusion. The pre-defined options are:
  - "The document did not show up";
  - "Focus words not visible";
  - "Poor text quality or composition";
  - "Language is too difficult";
  - "Refers to things in other documents";

  - "Too Long or too Short";
  - "Wrong sense of focus word";
  - "Not in Portuguese";
  - "Too much slang";
  - "Bad or Uninteresting Topic";
  - "Other (specify below)";

- Additional comments (optional) – where the teacher can give more clues about his/hers decision of exclusion.

With this expert information, REAP.PT can now prevent students from seeing the excluded documents, and give priority to documents that have been marked as "Excellent". Of course, the ratings are associated with a given teacher and it is possible to consider the ratings according to the class of each student (whether it was the teacher from the class the student is engaged in that ranked the document). This feature will make REAP.PT a platform that evolves, and improves (with expert knowledge), at the same time that is being used.

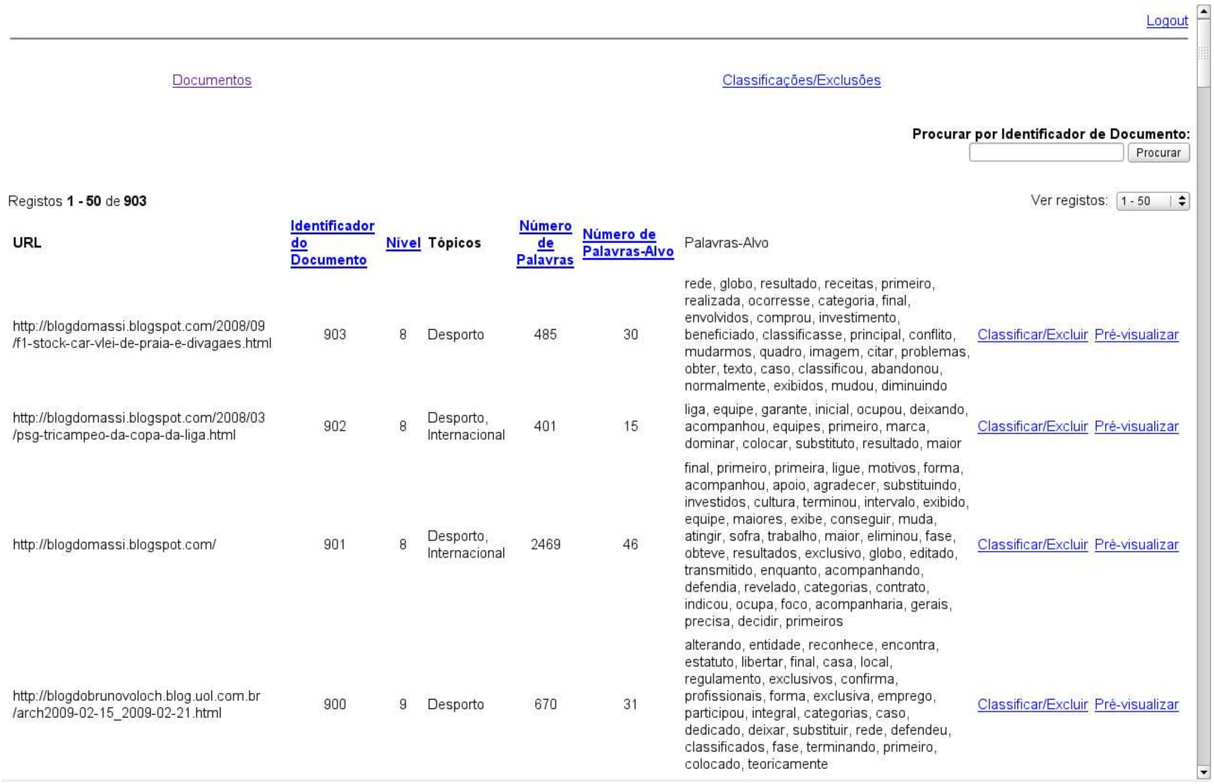
At last, on this same system, the teacher can view the ratings that have been assigned to the documents, being able to modify or even delete them. This is done using a menu (on the top of the page) where the user can choose between the document and the rating view. To better illustrate the information it provides and how it is provided, Figure 4.4 shows the interface of the *Document Viewing System*.

## 4.3 Speaker Diarization in the Broadcast News

One of the features the Portuguese version of REAP introduced was the *Oral Comprehension Module*. The motivation behind the development of this module was the scarce existence of materials that allow the students to hear someone speak in Portuguese. Although Portuguese is a language where the correspondence between graphemes and phones is in most cases regular, if one focus on the European variety of Portuguese (object of study of the REAP.PT system), the phenomenon of vowel reduction is always present, even in careful speech.

REAP.PT innovated by introducing a synthesizer for the oral reading of any selected text, by inserting an "Audio Book" section (where speech and text are aligned), and also by inserting a section of Broadcast News (where audio, video and text are aligned). The original version of these innovations was developed and integrated in the REAP.PT interface by José Lopes [24]. The contribution of the present work to this section is divided in two parts.

In the first place, there were two versions of the REAP.PT system that were being developed separately without the aid of a Revision Control System. This fact became alarming when the database



**Procurar por Identificador de Documento:**

Registos **1 - 50** de **903**

Ver registos: 1 - 50

| URL | Identificador do Documento | Nível | Tópicos | Número de Palavras | Número de Palavras-Alvo | Palavras-Alvo | | |
|---|---|---|---|---|---|---|---|---|
| http://blogdomassi.blogspot.com/2008/09/f1-stock-car-vlei-de-praia-e-divagaes.html | 903 | 8 | Desporto | 485 | 30 | rede, globo, resultado, receitas, primeiro, realizada, ocorresse, categoria, final, envolvidos, comprou, investimento, beneficiado, classificasse, principal, conflito, mudarmos, quadro, imagem, citar, problemas, obter, texto, caso, classificou, abandonou, normalmente, exibidos, mudou, diminuindo | Classificar/Excluir | Pré-visualizar |
| http://blogdomassi.blogspot.com/2008/03/psg-tricampeo-da-copa-da-liga.html | 902 | 8 | Desporto, Internacional | 401 | 15 | liga, equipe, garante, inicial, ocupou, deixando, acompanhou, equipes, primeiro, marca, dominar, colocar, substituto, resultado, maior | Classificar/Excluir | Pré-visualizar |
| http://blogdomassi.blogspot.com/ | 901 | 8 | Desporto, Internacional | 2469 | 46 | final, primeiro, primeira, ligue, motivos, forma, acompanhou, apoio, agradecer, substituindo, investidos, cultura, terminou, intervalo, exibido, equipe, maiores, exibe, conseguir, muda, atingir, sofra, trabalho, maior, eliminou, fase, obteve, resultados, exclusivo, globo, editado, transmitido, enquanto, acompanhando, defendia, revelado, categorias, contrato, indicou, ocupa, foco, acompanharia, gerais, precisa, decidir, primeiros | Classificar/Excluir | Pré-visualizar |
| http://blogdobrunovoloch.blog.uol.com.br/arch2009-02-15_2009-02-21.html | 900 | 9 | Desporto | 670 | 31 | alterando, entidade, reconhece, encontra, estatuto, libertar, final, casa, local, regulamento, exclusivos, confirma, profissionais, forma, exclusiva, emprego, participou, integral, categorias, caso, dedicado, deixar, substituir, rede, defendeu, classificados, fase, terminando, primeiro, colocado, teoricamente | Classificar/Excluir | Pré-visualizar |

Figure 4.4: *Document Viewing System* interface.

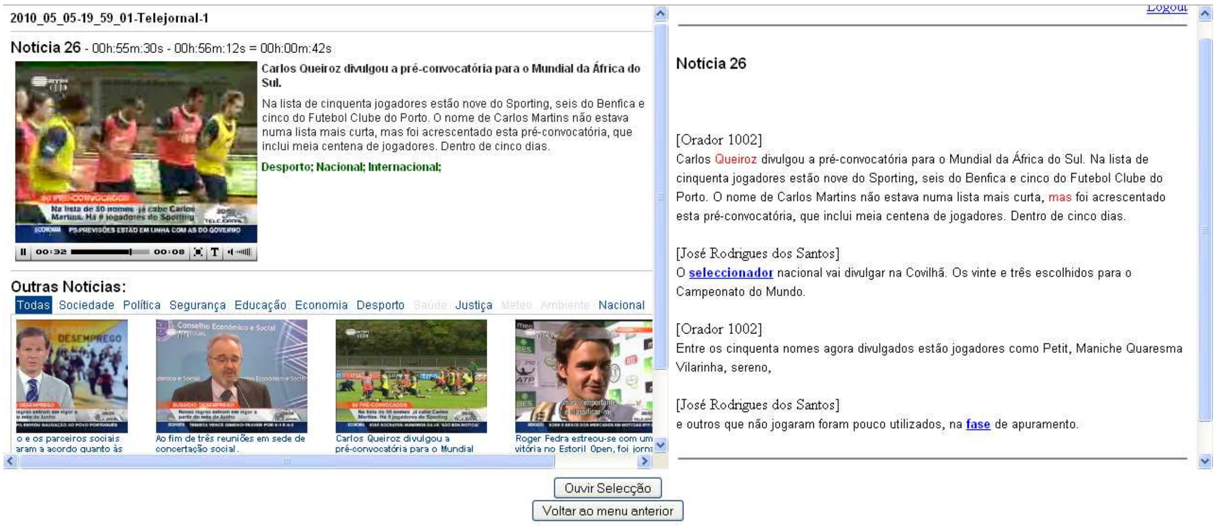
was changed, and the systems were now difficult to merge. For that reason, an effort was put into the integration of the two versions, adapting the needed parameters to have a coherent and fully working system, in spite of always having to open different pages to show different functionality. This new merged version was then submitted to the project SVN, that already existed but was not being used.

The second contribution regarding the *Broadcast News* topic was the development of a script to assign a speaker to the utterances that were recognized by the automatic speech recognition system (ASR). AUDIMUS [30] is the ASR system used in L$^2$F, being tuned for this task since it takes advantage of a large corpus of BN stories, manually transcribed, that was used for training and testing the system. Since the output of AUDIMUS already provides information about the speakers, the new insertion script collected that information and transformed the transcription of the BN story in a dialogue, showing the speaker that is talking at each time.

AUDIMUS additionally provides information on the gender of the speaker and, for the most common speakers (such as anchors, politicians, etc.), provides their name.

Since there are still errors with the gender distinction tool it was decided not to distinguish speakers based on their gender. On the other hand, the information about specific speakers is accurate and, for that reason, when a speaker is known, his/her name is presented to the student instead of simply its identifier.

Figure 4.5 presents the new interface of the Broadcast News, where the speakers' diarization can be clearly seen ont the right side of the page. Moreover, words that were recognized with low confidence ($< 82\%$) are marked in red, and *Focus Words* are highlighted in blue.



2010_05_05-19_59_01-Telejornal-1

**Notícia 26** - 00h:55m:30s - 00h:56m:12s = 00h:00m:42s

**Carlos Queiroz divulgou a pré-convocatória para o Mundial da África do Sul.**

Na lista de cinquenta jogadores estão nove do Sporting, seis do Benfica e cinco do Futebol Clube do Porto. O nome de Carlos Martins não estava numa lista mais curta, mas foi acrescentado esta pré-convocatória, que inclui meia centena de jogadores. Dentro de cinco dias.

**Desporto; Nacional; Internacional;**

Outras Notícias:

Todas Sociedade Política Segurança Educação Economia Desporto Saúde Justiça Meteo Ambiente Nacional

o e os parceiros sociais aram a acordo quanto às

Ao fim de três reuniões em sede de concertação social.

Carlos Queiroz divulgou a pré-convocatória para o Mundial

Roger Fedra estreou-se com um vitória no Estoril Open, foi jorn

Logout

**Notícia 26**

[Orador 1002]
Carlos Queiroz divulgou a pré-convocatória para o Mundial da África do Sul. Na lista de cinquenta jogadores estão nove do Sporting, seis do Benfica e cinco do Futebol Clube do Porto. O nome de Carlos Martins não estava numa lista mais curta, mas foi acrescentado esta pré-convocatória, que inclui meia centena de jogadores. Dentro de cinco dias.

[José Rodrigues dos Santos]
O **seleccionador** nacional vai divulgar na Covilhã. Os vinte e três escolhidos para o Campeonato do Mundo.

[Orador 1002]
Entre os cinquenta nomes agora divulgados estão jogadores como Petit, Maniche Quaresma Vilarinha, sereno,

[José Rodrigues dos Santos]
e outros que não jogaram foram pouco utilizados, na **fase** de apuramento.

Figure 4.5: Broadcast News interface.

# 5 Conclusion

This last chapter presents the final remarks of this thesis, summarizing the work that was accomplished. It concludes by presenting some ideas for future work.

## 5.1 Final Remarks

REAP.PT arrives at the current research panorama as an adaptation of a project in the CALL area (REAP) for European Portuguese. But, as expected and desirable, the porting tasks introduce new features (as the oral comprehension module), new problems and solutions. In fact, the expansion of this kind of system to different languages benefits from the idiosyncrasies of each language that gives rise to those new questions, answers and solutions. Of course, all languages can take advantage of these.

None of the developed types of questions can assure the third and last level of Stahl's theory. A question capable of assuring it would have to be complex, demanding a deep knowledge of the text subject and for free-form responses. At present, there are no Natural Language methods that could ensure the inclusion of these questions with enough quality to be considered.

This work provided several resources and mechanisms to REAP.PT that are important for future development. The construction of a coherent set of *focus words*, with readability level classification, related with the questions that aim to test them, are examples of the output of the present thesis.

This work also compared several ways of producing *distractors* for *cloze* questions. For native testees, the phonetic approach provided the highest correction rate, closely followed by the manually produced distractors and the graphemic plus filtering approach. The random (with and without filtering) and the graphemic approaches yielded the lowest results.

For non-native speakers, despite their low representativeness, the highest correction rate was obtained with the graphemic plus filtering approach, and the lowest one with the graphemic approach. The phonetic approach and the two approaches with filtering yielded correction rates which are similar to the one achieved by the manually generated *distractors*, indicating a good adequacy for the integration of these approaches in the REAP.PT system.

The use of different strategies for automatically generating distractors according to the student's level of comprehension of the language seems useful to guide the student throughout the vocabulary learning process of Portuguese. The false-friends *distractors*, used in the manual generation, are also an interesting topic for future research. Another one is the inclusion of *collocations*, thus being able to increase the difficulty and correctness of *distractors*, generating the ones that have low collocation values in a specific *stem*.

Regarding resources, unlike for English, it is difficult to find resources in European Portuguese that meet the needs for this task and that are actually reliable. The work on *cloze* question generation would significantly benefit from the integrating of new/larger resources, with much higher recall rates, and more exhaustive semantic information.

For Portuguese, one thus have a long work ahead of us.

## 5.2 Future Work

REAP.PT in general and the automatic question generation task in particular, can both benefit from further research in the area of Natural Language Processing. The following sections present a selection of topics that could improve the system and can be considered as future work.

### 5.2.1 Integration of syntactic information

With the inclusion of syntactic information in the readings that are presented to the student, REAP.PT would be capable of identify in a more informed way which sense of the word a particular occurrence relates to. Apart from this, syntactic information could also help in the selection of the documents, excluding texts that are syntactically badly formed.

### 5.2.2 Word Sense Disambiguation module

One of the goals of REAP.PT is to be able to identify the correct sense of a word in a given text. For example, when a student does a dictionary lookup, the sense of the word in the text should be highlighted in the definition text. This has been a hot topic in the last few years and is really a necessity in systems like REAP, that aim to teach a language and, therefore, demand for correctness.

### 5.2.3 *Stem* generation for *cloze questions*

The *stems* used in the *cloze* questions should be automatically generated. P-AWL and REAP scope can change anytime, and manual development of this resource is extremely fastidious and not error free. For these reasons, a specific module to generate this kind of information is necessary and should also be able to suppress some failures, taking into account, systematically, aspects like the length of the sentences, their level, etc.

### 5.2.4 Syntactic exercises

So far, REAP.PT is endowed with mechanisms that provide questions about vocabulary. Vocabulary is part of a language but, to be considered proficient there are rules that need to be followed. The development of exercises that are able to test other aspects of the language, as grammar, would be an interesting way to complement the system. Actually, this work is currently being investigated by Cristiano Marques in his Master thesis work.

### 5.2.5 Games

Carnegie Mellon University started to develop games for inclusion in the REAP system, a work that has been carried out by Adam Skory. The main goal is to make the learning process more attractive and appealing for the students. André Silva is researching the subject, and proposes to endow REAP with pictorial games in his Master thesis.

### 5.2.6 Interface and Portability

The simplicity of the current REAP.PT interface may not be an attractive point to some students. For that reason, REAP.PT needs to meet the students' interests and be adapted to new technologies – devices that the students use in their everyday life. João Sirgado, in his Master thesis, is researching methods of porting REAP.PT to portable devices (as iPad and iPhone).

### 5.2.7 Classify Documents Regarding Portuguese Varieties and Type

The method used in this thesis to restrict the variety of Portuguese (whether Brazilian or European) simply relies on the description of the URL. The same technique is used to exclude documents that are expected to be *blog* content. One should notice that the fact that a web page being in a *.PT* domain does not mean that it is written in European Portuguese. For that reason, the inclusion of these new filters in the chain is left as a proposed future work.

# Bibliography

[1] Luís Marujo. REAP.PT. M.Sc. Dissertation in Information Systems and Computer Engineering, Instituto Superior Técnico - Universidade Técnica de Lisboa, 2009.

[2] Rui Correia, Jorge Baptista, Nuno Mamede, Isabel Trancoso, and Maxine Eskenazi. Automatic Generation of Cloze Question Distractors. In *Second Language Studies: Acquisition, Learning, Education and Technology*, Tokyo, Japan, September 2010.

[3] Joy L. Egbert and Gina Mikel Petrie. *Call research perspectives*. ESL and Applied Linguistics Professional Series. Lawrence Erlbaum Associates, Mahwah, 31 May 2005.

[4] Michael Heilman, Kevyn Collins-Thompson, Jamie Callan, and Maxine Eskenazi. Classroom success of an Intelligent Tutoring System for lexical practice and reading comprehension. In *Ninth International Conference on Spoken Language Processing*. Citeseer, 2006.

[5] Kevyn Collins-Thompson and Jamie Callan. Information retrieval for language tutoring: An overview of the REAP project. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 544–545, New York, NY, USA, 2004. ACM.

[6] Jorge Baptista, Neuza Costa, Joaquim Guerra, Marcos Zampieri, Maria de Lurdes Cabral, and Nuno Mamede. P-AWL: Academic Word List for Portuguese. *Computational Processing of the Portuguese Language*, pages 120–123, 2010.

[7] Jorge Baptista. Portuguese Academic Word List (P-AWL). v0.1, 15 June 2009.

[8] Rui Amaral, Hugo Meinedo, Diamantino Caseiro, Isabel Trancoso, and João Neto. A Prototype System for Selective Dissemination of Broadcast News in European Portuguese. *EURASIP journal on Advances in Signal Processing*, 2007(24):1–12, 2007.

[9] Luís Oliveira, Céu Viana, and Isabel Trancoso. DIXI – A Generic Text-to-Speech System for European Portuguese. In *PROPOR'2008 - 8th International Workshop on Computational Processing of the Portuguese Language*, Aveiro, Portugal, 2008. Springer-Verlag.

[10] *Diário da República*. Number 193, I-A. 23 August 1991.

[11] Della Summers, editor. *Longman Active Study Dictionary*. Addison Wesley Longman, 1998.

[12] Steven A. Stahl. Three Principles of Effective Vocabulary Instruction. *Journal of Reading*, 29(7):662–668, 1986.

[13] Christine M. Feeney and Michael Heilman. Automatically Generating and Validating Reading-Check Questions. In *ITS '08: Proceedings of the 9th international conference on Intelligent Tutoring Systems*, pages 659–661, Berlin, Heidelberg, 2008. Springer.

[14] Jonathan C. Brown, Gwen A. Frishkoff, and Maxine Eskenazi. Automatically Question Generation for Vocabulary Assessment. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 819–826. Association for Computational Linguistics, 2005.

[15] Robert Gallisson and Daniel Coste. *Dicionário de Didáctica das Línguas*. Livraria Almedina, Coimbra, 1983.

[16] Wolfgang Köhler. *Gestalt Psychology*. Liveright, New York, 1947.

[17] Juan Pino, Michael Heilman, and Maxine Eskenazi. A Selection Strategy to Improve Cloze Question Quality. In *Proceedings of the Workshop on Intelligent Tutoring Systems for Ill-Defined Domains. 9th International Conference on Intelligent Tutoring Systems, Montreal, Canada*, pages 22–32. Citeseer, 2008.

[18] Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1 edition, 18 June 1999.

[19] Arthur C. Graesser and Robert A. Wisher. Question Generation as a Learning Multiplier in Distributed Learning Environments, 2001.

[20] David Coniam. A Preliminary Inquiry into Using Corpus Word Frequency Data in the Automatic Generation of English Language Cloze Tests. *CALICO Journal*, 14(2):15–34, 1997.

[21] Siddharth Patwardhan and Ted Pedersen. Using WordNet-based Context Vectors to Estimate the Semantic Relatedness of Concepts. In *Proceedings of the EACL 2006 Workshop Making Sense of Sense - Bringing Computational Linguistics and Psycholinguistics Together*, pages 1–8, Trento, Italy, 2006.

[22] Hubbard C. Goodrich. Distractor Efficiency in Foreign Language Testing. *TESOL Quarterly*, 11(1):69–78, 1977.

[23] Juan Pino and Maxine Eskenazi. Semi-Automatic Generation of Cloze Question Distractors Effect of Students' L1. In *Proceedings of the 2009 Workshop on Speech and Language Technologies in Education (SLaTE), Wroxall Abbey Estates, United Kingdom*, 2009.

[24] Luís Marujo, Nuno Mamede, José Lopes, Isabel Trancoso, Juan Pino, Maxine Eskenazi, Jorge Baptista, and Céu Viana. Porting REAP to European Portuguese. In *ISCA International Workshop on*

*Speech and Language Technology in Education (SLaTE 2009), ISCA, Wroxall Abbey Estate, Warwickshire*. Citeseer, September 2009.

[25] V. Levenshteiti. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet Physics-Doklady*, volume 10, 1966.

[26] Daniel Gomes and Mário Silva. The Viúva Negra Crawler: an experience report. *Software: Practice and Experience*, 2(38):161–168, 2008.

[27] ISO 28500:2009. Information and documentation — The WARC File Format. June 2009.

[28] Christian Kohlschütter, Peter Fankhauser, and Wolfgang Nejdl. Boilerplate Detection using Shallow Text Features. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 441–450. ACM, 2010.

[29] Jeffrey Dean and Sanjay Ghemawat. MapReduce: Simplified Data Processing on Large Clusters. *Communications of the ACM*, 51(1):107–113, 2008.

[30] Hugo Meinedo, Márcio Viveiros, and João Neto. Evaluation of a Live Broadcast News Subtitling System for Portuguese. In *Proc of Interspeech*, 2008.

# I

# Appendices

# A

# P-AWL Words by Level

## A.1 5th Grade

- abandonar
- academia
- acumulação
- acumular
- adaptar
- adulto
- afectivamente
- afixar
- ajuda
- ajudar
- ajuste
- anual
- apoio
- área
- assegurar
- assunção
- assunto
- atribuição
- aumentar
- avaliar
- benefício
- bolsa
- canal
- canalizado
- canalizar
- capaz
- capítulo
- característico
- categoria
- chamado
- ciclo
- citação
- clareza
- colapso
- colega
- colocar
- começar
- começo
- compor
- compra
- comprador
- compreender
- computador
- comunicar
- concluir
- confiança
- confiar
- conformar
- conforme
- consagrar
- conseguir
- constantemente
- consulta
- consultar
- contacto
- contratar
- convencido
- conversar
- converso
- corporação
- debate
- decidir
- declinar
- deduzir
- definido
- deixar
- depressão
- deprimir
- desafio
- descobrir
- desmotivar
- detector
- dominar
- empresa
- encontrar
- enorme
- equivaler
- errar
- esclarecer
- esclarecimento
- esquema
- essência
- estabelecimento
- estimar
- estudar
- estudo
- eticamente
- eventualidade
- expansão
- exploração
- exteriorizar
- extrair
- fácil
- finalmente
- física
- forçar
- formato
- formular
- funcionamento
- fundação

fundar
fundo
gráfico
herança
ignorante
ilegalidade
imaginar
imaginativo
imaturo
incessante
inconformado
indiscreto
individualizar
inevitável
informação
informático
insignificante
inspeccionar
instrutivo
instrutor
inteligente
intensamente
intervalo
inútil
inutilmente
inválido
juntar
labor
legislação
libertar
licenciado
localizar
maduro
manual
mapa
marcar
margem
máximo
médico
mental
meta
ministro
modalidade
monitor
montar
negar
negativa
número
obter
ocupação
orientar
origem
palestra
par
parágrafo
parcialmente
participar
participativo
período
perseguição
perseguir
persistente
polémica
posar
preciso
princípio
problema
procurar
profissão
profissional
projectar
propriedade
racional
rascunho
reacção
realizar
receita
recuperar
recurso
rede
regime
rejeitar
renda
repartir
requisitar
resolver
responder
resposta
resumir
retenção
revelação
rever
rota
salientar
secção
segurança
sexo
significado
somar
subsídio
sucesso
suposição
tabela
tarefa
técnica
técnico
temática
terreno
texto
total
trabalho
transporte
unicamente
útil
variar
vestígio
visivelmente

## A.2 6th Grade

acrescentar
adjacente
administrar
afectar
agradecer
alterado
alvo
analisar
análogo
aparentemente
apor
apreciar
apropriado
aproximar
assim
auxílio
avaliação
calcular
central
citar
classificar
colaborar
comentar
comentário
complemento
composto
conceber
conceder
concentrado
conclusão
conclusivo
conflito
consentir
constante
constituição
construção
construir
construtor
contemporâneo
contradição
contradizer
contrário
contrato
converter
convocar
cooperar
criação
cultura
dado
decorrer
defender
defensor
demonstração
demonstrar
deprimido
derivar
desabar
descartar
desencadear
deslocar
directo
distintamente
distribuição
distribuir
diversificar
documentar
domesticado
domesticar
doméstico
ecológico
economia
editora
editorial
eliminar
emergir
emigrante
encontro
envolver
equipa
escala
esférico
estilista
estilo
estrada
evidente
evolução
exactamente
exibição
expor
exposição
facilidade
facilitar
feito
fenomenal
ferido
finalizar
flutuante
forma
garantir
idêntico
identificação
ignorância
ignorar
ilegal
ilustração
ilustrar
implicar
impropriamente
incapaz
indicação
indicativo
indiscrição
individualmente
infinito
informar
inovador
inovar
inspector
instituto
insuficiente
inteligência
interno
intérprete
invariável
investigar

involuntário
irregular
libertação
liberto
ligação
ligar
localização
maior
melhorar
método
migratório
militar
missão
modo
moeda
motivar
mudar
negação
neutralizar
obrigação
ocupar
orientação
oriental
oriente
painel
parâmetro
passivo
periódico
pesquisa
pesquisar
porção
praticante
precedente
precisar
predizer
presumido
prever
previsão
primário
primeiro
principalmente
procedimento
proporcionado
prosseguir
publicação
qualidade
química
reagir
realizável
recomeçar
recomeço
reconhecer
reforçar
região
reiniciar
residência
residir
restringir
resultado
resumo
reter
reverso
revisão
rigidez
saída
seguro
série
significar
simulação
sítio
sociedade
sofrer
soma
submeter
subsidiar
suceder
suficiente
suplementar
suspender
temporariamente
terminar
traçar
traço
tradicionalista
transformação
transição
transmissão
transportar
último
único
uniforme
uniformidade
unir
variável
veículo
via
volume
voluntário

## A.3 7th Grade

abandonado

acção

accionar

acompanhado

acompanhar

adaptação

adequação

adequadamente

adequado

adequar

admitir

adquirir

advogado

agenda

ajustar

alcançar

algo

alojamento

alteração

alterar

alternativa

alvejar

ambiente

análise

analítico

anexo

anormal

antecipar

anualmente

aparentar

apêndice

apreciado

aproximação

aproximadamente

aproximado

aspecto

assembleia

assistência

assistir

assumir

atingir

atitude

atribuído

atribuir

atributo

automático

autómato

autor

autoridade

auxiliar

breve

camada

capacidade

característica

caracterização

caracterizar

casal

casar

cenário

centro

cerne

cessar

circunstância

claridade

clássico

classificação

classificado

coerente

coincidência

coincidir

compatível

complexo

componente

comprar

compras

computadorizado

comunicação

comunicativo

comunidade

concentração

concentrar

conduzir

conferir

confirmar

conflituoso

conformação

consciência

consequência

consequentemente

considerável

consistência

consistir

constituinte

constituir

constitutivo

consultório

consumir

contactar

conter

contexto

contraditório

contrariamente

contraste

contribuição

contribuir

convencer

convincente

coordenação

coordenado

correspondência

corresponder

criador

criar

critério

culto

cultural

debater

década

decisão

dedicar

definição

definir

demonstrativo

denunciar

derivação

derradeiro
desafiar
desajustado
desfecho
deslocação
desregular
destacar
destaque
detectar
detective
diferenciar
diminuir
discretamente
discreto
discrição
disponível
distinção
distinto
distorcer
diverso
diversos
documentação
dominante
domínio
dramático
dramatização
dramaturgo
duração
ecologia
ecologista
edição
editar
elemento
emigrar
empreendedor
emprego
energia
enquadramento
enquanto
entidade
envolvido
equipamento
equivalente
erro
especificar
específico
esquematizar
essencial
estabelecer
estatístico
estável
estratégia
estratégico
estrutura
estruturar
etiqueta
evidência
evidenciar
evidentemente
evitar
exactidão
exacto
exclusivamente
exclusivo
expandir
explorar
exterior
externo
factor
fase
fenómeno
ferir
filme
filosofia
final
finalidade
finanças
fisicamente
físico
fita
flexível
flutuar
focar
fonte
formal
fórmula
função
funcionar
fundamental
fundamentalmente
género
geração
geral
gerar
hierárquico
hipótese
identidade
identificar
identificável
imagem
imaginário
implicitamente
implícito
impor
imposto
impreciso
imprevisível
impróprio
inadequadamente
incidente
inclinação
inclinar
incluir
inconsciente
indefinido
indicador
indicar
índice
indivíduo
inédito
infinitamente
informal
inibidor
iniciado
inicial
inicialmente
iniciar
iniciativa
início
inquérito
inseguro
inserir
inspecção
instituição
instrução

integral
integrar
íntegro
intensidade
intensificar
interacção
intermédio
interpretar
intervenção
interveniente
intervir
inversamente
inverter
investigador
invisível
invocar
involuntariamente
isolar
item
jornal
justificação
justificar
liberdade
licença
licenciar
ligado
limitar
local
lógico
maioria
manter
marca
mecanismo
meio
menor
mentalmente
mínimo
ministerial
ministério
modificação
modificar
mostra
motivo
mutuamente
negativo
noção
norma
normal
normalmente
núcleo
objectivo
objecto
obviamente
óbvio
ocorrência
ocorrer
opção
opcional
oração
papel
parcial
participação
participante
passivamente
património
perceber
peritar
permitir
perspectiva
perspicácia
política
político
pose
positivo
preceder
precisamente
precisão
preconceito
predominante
predominantemente
predominar
presunçoso
principal
proceder
processo
proibir
projectista
projecto
proporção
proporcional
próximo
psicológico
publicar
quadro
questão
quota
radical
realização
reconstruir
recriar
refinamento
reforço
regional
registar
registo
regular
relacionar
relatar
relatório
relutância
requerer
requisito
residente
ressaltar
restaurar
resumidamente
revelar
revisor
revista
revolução
rígido
rotular
selecção
seleccionar
sequência
sexual
simbólico
simbolizar
símbolo
simular
síntese
sintético
site
sobreposição
sobreviver

sublinhar

subordinação

subordinar

substituição

substituir

sucessão

sucessivo

sucessor

suficientemente

supor

suspensão

sustentar

sustento

teatral

tecnológico

tema

temático

tendência

tensão

título

todavia

tradição

tradicional

transformar

trânsito

transmitir

união

universidade

utilizar

válido

vantagem

variação

versão

visível

visual

## A.4 8th Grade

abandono
abordado
abordar
académico
acesso
acomodação
acomodar
acompanhamento
acumulado
administração
administrativo
advocacia
afectado
afectivo
afecto
agradecimento
alcance
ambiental
ambientalista
aproveitamento
aquisição
assistente
aumento
automaticamente
autoritário
brevemente
brevidade
centrar
civil
clarificar
código
colaborador
comissão
comissário
comodidade
compensação
comprometer
compromisso
comunicado
conferência
congresso
consciencializar
consciente
consenso
constitucionalmente
constranger
consumidor
consumo
contactável
contrastar
convenção
conversação
convir
cooperação
cooperativa
coordenar
cotar
criatividade
criativo
dedução
definitivamente
definitivo
denotar
desacompanhado
desafiador
design
desintegração
desintegrar
desligado
desperto
desviar
dimensão
diminuição
dinâmico
discriminação
discriminar
distribuidor
divulgação
económico
editor
eliminação
elo
emergência
emparelhar
empreender
empresarial
ênfase
envolvimento
equação
equipar
esfera
esquematicamente
estatística
estatuto
estimativa
etnia
eventual
evoluir
excessivo
excesso
excluir
exibir
exportador
exteriorização
federação
federal
ferimento
filosófico
financeiramente
financiar
foco
frisar
garante
global
globo
inabordável
inalterado
inalterável
incoerente
inconcebível
inconsciência
inconstante
incorporação
indistintamente
individualismo

induzir
inexacto
informática
inovação
insegurança
instabilidade
instável
instruir
insuficientemente
insustentável
intensivo
interpretação
investigação
investir
irracional
irreversível
isolamento
legislar
lesão
major
marginal
media
minimalista
montagem
mundialmente
mútuo
nuclear

paralelo
parceiro
perseguidor
persistir
perspicaz
precedência
predominância
pressuposto
previamente
prévio
previsto
prioritariamente
prioritário
problemática
promover
prospectar
psicologicamente
químico
reactor
recentrar
reconhecimento
refinado
reformular
registrar
regulamentar
rejeição
relaxar

relevar
resolução
rótulo
sector
seguramente
significativo
similar
sintetizar
sobrepor
sobrevivência
sobrevivente
status
subjectivamente
subjectividade
sucessivamente
sumário
sumarizar
suplemento
tecnologia
tenso
teoria
tese
textual
textualmente
tópico
transferir
unificar

universitário
utilizador
variante
violar
virtual
visão
visualização
visualmente
volumoso

## A.5 9th Grade

abordagem
acessível
acompanhante
agregar
aleatório
aparente
apreciável
arbitrário
atribuível
beneficiar
beneficiário
compensar
complementar
comprometido
conceito
concepção
condução
conduta
conformidade
conformismo
consentimento
contextualizar
convencional
conversão
convocação
denúncia
desajustamento
desocupar
desproporção
desvio
detectável
devoção
devotar
dinâmica
distintivo
drama
economicamente
enormemente
erosão
étnico
evolutivo
exceder
exclusão
expansionismo
extracção
extracto
financeiro
flutuação
funcionalmente
garantia
globalmente
hierarquia
ideológico
imparcial
imparcialidade
imparcialmente
implementar
imposição
imprevisto
incidir
inconstitucional
indemnização
inevitavelmente
informatizado
inibição
iniciação
instituir
integralmente
integridade
intensificação
intrínseco
inviolável
invisivelmente
irrelevante
justificado
legal
legalidade
legalmente
liberal
liberar
maturo
melhoria
mentalidade
minimalismo
minoria
motivação
nega
negativamente
obtenção
paralelismo
parcialidade
percentagem
persistência
preliminar
presumir
presunção
proibição
projecção
promoção
propina
psicologia
racionalidade
recriação
refrear
reinvestir
relevante
remoção
remover
rendimento
requisição
rigidamente
sensivelmente
simbolicamente
simbolismo
somatório
subjectivo
subordinado
substituto
suficiência
sumariamente
sumariar
transitar
turno
ultramarino
utilmente
visualizar

## A.6 10th Grade

abrangente
abstracção
abstrair
adaptável
administrador
admissão
agendar
agregado
aleatoriamente
alternativo
amadurecer
ambiguidade
ambíguo
analisador
antecipação
antecipado
arquivar
arquivo
benéfico
capacitar
clarificação
cláusula
codificar
coerência
coerentemente
compilação
complexidade
comprometimento
computação
concebível
conceptual
conceptualmente
concessão
conformado
consensual
consideravelmente
consistente
constrangedor
constrangimento
consultor
contenção
contextual
contratante
contribuinte
controverso
crédito
culturalmente
declínio
decorrente
dedicadamente
deformar
desconstrutivo
desdramatizar
desestruturação
deslocamento
despromover
diferenciação
disponibilidade
disponibilizar
dispositivo
distorção
distribuidora
diversidade
documental
elementar
empírico
empreendimento
energético
equacionar
equivalência
especificação
especificamente
especificidade
esquemático
estabilidade
estabilizar
estilizar
estruturação
estruturalmente
ética
eventualmente
evitável
exibicionista
explicitamente
explorador
exportar
facilidades
facilitação
filosofar
finalização
financiamento
finito
flexibilidade
formalização
formalizar
formalmente
formatação
formulação
funcional
generalizado
globalização
hipotético
ideologia
ilógico
imigrante
impacto
ímpar
implementação
implicação
inacessível
inadequação
inatingível
incapacitar
incentivo
incidência
incoerência
incompatível
inconsistente
inconstância
incorporar
indefinidamente
indefinível
indemne
indemnizar
inegável

inerente
inferir
inserção
insignificância
institucional
institucionalizar
insuficiência
integração
intensão
interactivo
interagir
internamente
intuitivo
invariavelmente
inversão
investidor
invisibilidade
justificadamente
laboral
lesionar
logicamente
manutenção
maximizar
menoridade
metódico
metodologia
migrar
minimizar
ministrar
modal
monitorizar
neutral
normalidade
normalização
nortear
objectivamente
objectividade
ocupante
offset
parceria
periodicamente
perito
perspectivar
politicamente
potencial
potencialmente
pressupor
previsibilidade
previsível
primazia
prioridade
processar
profissionalmente
proporcionalmente
prospecção
protocolo
psicólogo
qualitativo
radicalmente
reagente
reconstrução
recuperação
redefinição
redefinir
redistribuir
refinar
reformulação
regulador
regulamento
relaxamento
relevância
relutante
requerente
requerimento
restrição
restritivo
reversível
reverter
revolucionar
revolucionário
seleccionador
selectivo
sequencial
sequencialmente
sequenciar
significante
significativamente
sobrevida
stress
stressar
subestimar
subsidiariamente
suposto
sustentável
tecnicamente
tematicamente
temporário
teoricamente
terminal
tradicionalmente
transferência
transferível
transitório
transportador
unificação
utilitário
validade
variavelmente
vínculo
violação
violador
virtualmente
visibilidade
voluntariado

## A.7 11th Grade

acessar
adaptabilidade
adaptativo
agregação
ajustado
ajustamento
alternativamente
analista
analiticamente
analogia
anexado
anexar
anormalidade
apreço
apropriadamente
avaliativo
canalização
cíclico
coincidente
comentador
compilador
compilar
complexificar
computacional
conceptualizar
conclusivamente
conferencista
confiadamente
confinante
confirmação
constitucional
construtivo
consultoria
contrastante
contrastivo
contratação
convencionalmente
dedutivo
deformação
demonstrável
denotação
deprimente
desadequado
descoordenação
desestabilizar
desestruturar
desfocar
designer
desmotivação
desproporcionado
detecção
devotadamente
diminuído
dinamismo
dinamizar
directriz
disseminado
diversamente
diversificação
dominação
ecologicamente
economista
emergente
emigração
empiricamente
enfático
enfatizar
enormidade
erroneamente
erróneo
estabilização
estrategicamente
estrutural
etiquetar
expansivo
exportação
expositor
facilitador
filosoficamente
finança
garantidamente
geracional
ideologicamente
ilustrativo
imigração
incompatibilidade
inconformidade
inconsequência
inculto
indiscretamente
individualista
inerência
inflexível
inibir
inigualável
injustificado
instância
institucionalização
interventivo
intrinsecamente
intuição
invariante
investimento
legislador
legislativo
liberalismo
maioridade
manipulação
manipular
maturidade
mediação
metodológico
minimal
minimamente
neutralização
neutro
normalizar
paradigma
percepção
persistentemente
positivamente
predomínio
presumível
promocional
prospectivo

qualitativamente

racionalismo

racionalização

reaccionário

reactivação

reajustar

reajuste

reestruturação

refinação

reorientação

repto

residencial

restauração

restaurador

restauro

sexualmente

subjacente

subjazer

submissão

subsequentemente

subsidiário

suspensivo

taxa

teatralização

teórico

teorizar

transmissor

unicidade

uniformemente

validar

## A.8 12th Grade

acessibilidade
acomodado
acomodamento
advogar
ambiguamente
arbitrariamente
arbitrariedade
cessante
ciente
colapsar
compatibilizar
confinar
conformista
constância
controvérsia
cooperativo
coordenador
corporativo
cotação
credor
crucial
definível
denunciante
desabamento
desconstrução
desconstruir
descontextualizar
descoordenado
deslocalizar
despoletar
dominador
dramaticamente
empirismo
equiparar
escopo
estratega
estruturante
ético
expansionista
externamente
faseado
formatar
identicamente
imprecisão
imprevisibilidade
incomunicável
inconformismo
inconformista
indefinição
indistinto
indução
inevitabilidade
inferência
infundado
iniciador
interpretativo
intransferível
intuir
inutilizar
invalidar
involução
involutivo
irracionalidade
liberalização
libertador
manipulador
maturação
mediar
migração
neutralidade
obviar
paradigmático
passividade
pesquisador
preconceituoso
predição
presumivelmente
previsivelmente
problematizar
profissionalismo
promotor
prospectivamente
protocolar
racionalista
racionalizar
racionalmente
reavaliação
reavaliar
recuperável
regulação
reinterpretação
reinterpretar
revelador
sexualidade
simbolista
sinteticamente
subsequente
sustentabilidade
tendencioso
validação
variabilidade
voluntariamente

## A.9 No Level Assigned

abordável
academicamente
acessado
acoplado
administrativamente
aleatoriedade
alocação
alocar
alterável
ambientalmente
analogamente
anormalmente
apenso
apreciavelmente
asseguradamente
atingível
automação
automatizar
avaliador
avaliável
cartografado
categorização
ciclismo
codificação
coincidentemente
comissionado
comissionamento
compatibilidade
compensatório
complexar
computável
comunicável
conceitual
conceptualização
concomitante
concomitantemente
confiável
constituência
constructo
consultivo
consumpção
conversamente
conversível
convertível
cooperante
cooperativamente
creditação
creditar
criativamente
cronograma
crucialmente
debatível
decisor
demonstradamente
demonstrador
derivativo
desadequadamente
desafiante
desalojar
desambiguar
desanexado
desanexar
desapreciar
desapropriadamente
desapropriado
desestabilização
desformatação
desformatar
desocupação
despromoção
desproporcional
desproporcionalmente
desregulação
desregulador
desregulamentar
dimensional
dinamicamente
dissimilar
dissimilaridade
distribucional
distributivo
diversificadamente
eliminatória
empreendedorismo
enfaticamente
enrijecer
equipe
erodir
especificadamente
especificador
esquisso
estatisticamente
estilizadamente
estiloso
etiquetagem
etnicidade
evolucionista
externalizar
extractor
factorização

factorizar
fasear
fenomenologia
financiador
formalizável
hipoteticamente
honorários
ilegalmente
ilogicamente
imaturidade
imprecisamente
imprevisivelmente
inapropriadamente
inapropriado
incapacitado
incoerentemente
incompatibilizar
inconcebivelmente
inconclusivamente
inconclusivo
inconformar
inconforme
inconsistência
incontactável
incontroverso
inconvencional
indemnizatório
indexação
indexar
indisponibilidade
indisponível
inerentemente
inexactidão
inflexibilidade
informalmente
iniciante
injustificadamente
inovativo
input
instabilizar
instanciar
institucionalmente
instrutivamente
insustentabilidade
intensivamente
interactivamente
internalizar
invalidação
invariabilidade
invariância
inversor
involuir
irracionalmente
irrelevância
isolacionismo
isolacionista
legalista
legislatura
liberalista
liberalizar
licenciamento
maduramente
manipulativo
manualmente
mapear
marginalmente
maximização
medicamente
metodologicamente
minimalização
minimalizar
minimização
ministeriável
monitorização
motivadamente
multidimensional
nocional
obtível
obviedade
ocupacional
palestrante
paradigmaticamente
parametrizar
parametrizável
percentual
peritagem
porcento
preliminarmente
primariamente
proibitivo
protocolarmente
quotização
rácio
rastrear
rastreável
rastreio
reactivar
reactivo
reagendar
reajustado
reajustamento
reatribuir
reconstrutivo
reconstrutor
redefinível
redistribuição
reestruturar
regionalmente

reinício

reinvestimento

relaxante

relocalização

relocalizar

relutantemente

removível

reocorrer

reorientar

respondente

retransmissão

retransmissor

retransmitir

revisar

revisionismo

revisionista

rotulagem

seccionamento

seccionar

selectivamente

selector

sequenciamento

sexismo

sexista

significância

significantemente

similaridade

simulador

sobrestimar

stressante

substitutivo

subvenção

sumarização

suspensor

tecnologicamente

tendenciosamente

tópica

validamente

variância

vestigial

violável

voluntariar