



TÉCNICO LISBOA

Syntactic REAP.PT

Exercises on Clitic Pronouncing

Tiago Esteves de Freitas

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

Jury

President:	Professor Pedro Manuel Moreira Vaz Antunes de Sousa
Advisor:	Professor Nuno João Neves Mamede
Co-advisor:	Professor Jorge Manuel Evangelista Baptista
Evaluation Jury:	Professor Bruno Emanuel da Graça Martins

October 2012

Acknowledgements

I would like to thank my advisors, Professor Nuno Mamede and Professor Jorge Baptista for their guidance, suggestions and motivation.

I thank all my colleagues at Instituto Superior Técnico for their companionship and support throughout my degree and dissertation, namely to Rui Correia, Teresa Gama, Mário Almeida, José Lourenço, Andreia Guerreiro, Pedro Patrão, Ricardo Pires, Ricardo Sousa, and all others know who you are.

I would also like to thank my friends and family for their support and motivation at all times, with special thanks to my parents and my friends Yocelyn Correia, Raquel Marques, Luís Machado, João Lopes, Gonçalo Braz, João Pinheiro, José Guilherme, Luís Campos, Álvaro Meneses, and Diogo Vasconcelos.

Lisboa, November 11, 2012

Tiago Esteves de Freitas

To my parents and significant other.

Resumo

A investigação interdisciplinar em Aprendizagem Inteligente de Línguas Assistida por Computador (ICALL) visa integrar o conhecimento de linguística computacional com a aprendizagem de línguas assistida por computador (CALL). O REAP.PT é um projeto emergente nesta área, visando ensinar Português de forma inovadora e apelativa, adaptada a cada aluno. Este trabalho tem como objetivo melhorar o sistema REAP.PT, desenvolvendo novos exercícios sintáticos, gerados automaticamente, para ensinar o complexo fenómeno da pronominalização, isto é, a substituição de um constituinte por uma forma pronominal adequada. Embora esta transformação possa parecer simples, envolve complexas restrições lexicais, sintáticas e semânticas. Os problemas da pronominalização em Português tornam-na um aspecto particularmente difícil da aprendizagem da língua para falantes não-nativos. Mesmo os falantes nativos têm muitas vezes dúvidas quanto ao correto posicionamento dos clíticos, devido à complexidade e interação de fatores concorrentes que regem esse fenómeno. Uma nova arquitetura para a geração automática de exercícios sintáticos é aqui proposta. A mesma provou ser fundamental para o desenvolvimento deste exercício complexo e é esperado que constitua um contributo relevante na elaboração de futuros exercícios sintáticos, tornando-se, potencialmente, uma framework de geração automática deste tipo de exercícios. Também é aqui apresentado um sistema de feedback pioneiro, com explicações detalhadas e geradas automaticamente para cada resposta, e que permite melhorar a experiência de aprendizagem, como foi comentado pelos utilizadores. A avaliação de especialistas e utilizadores teve resultados positivos, demonstrando a validade da abordagem apresentada.

Abstract

The emerging interdisciplinary field of Intelligent Computer Assisted Language Learning (ICALL) aims to integrate the knowledge from computational linguistics into computer-assisted language learning (CALL). REAP.PT is a project emerging from this new field, aiming to teach Portuguese in an innovative and appealing way, and adapted to each student. The aim of this work is to improve the REAP.PT system, developing new, automatically generated, syntactic exercises. These exercises deal with the complex phenomenon of pronominalization, that is, the substitution of a syntactic constituent with an adequate pronominal form. Though the transformation may seem simple, it involves complex lexical, syntactical and semantic constraints. The issues on pronominalization in Portuguese make it a particularly difficult aspect of language learning for non-native speakers. On the other hand, even native speakers can be often uncertain of correct clitic positioning, due to the complexity and interaction of competing factors governing this phenomenon. A new architecture for automatic syntactic exercise generation is proposed. It proved invaluable in easing the development of this complex exercise, and is expected to make a relevant step forward in the development of future syntactic exercises, with the potential of becoming a syntactic exercise generation framework. A pioneer feedback system with detailed and automatically generated explanations for each answer is also presented, improving the learning experience, as stated in user comments. The expert evaluation and crowd-sourced testing results were positive, demonstrating the validity of the presented approach.

Palavras Chave Keywords

Palavras Chave

Ensino da Língua Assistido por Computador

Geração Automática de Exercícios

Exercícios Sintáticos

Língua Portuguesa

Keywords

Intelligent Computer Assisted Language Learning

Automatic Exercise Generation

Syntactic Exercises

Portuguese

Contents

1	Introduction	1
1.1	Goals	2
1.2	Document Structure	2
2	State of the Art	3
2.1	REAP.PT	3
2.1.1	REAP.PT Architecture	3
2.1.2	REAP.PT Exercises	5
2.1.3	REAP.PT Syntactic Exercises Architecture	8
2.2	Portuguese CALL Systems	10
2.2.1	Ciberescola	10
2.2.2	Aprender Português	12
2.3	ICALL Systems	12
2.3.1	TAGARELA	13
2.3.2	Working With English Real-Texts	13
2.3.3	The Alpheios Project	14
2.3.4	FAST	14
2.3.5	Arikiturri	16
2.4	Current Syntactic Exercises on Pronominalization	17
2.4.0.1	Common Student Errors	19

3	Exercise Generation Architecture	23
3.1	Rule Engine	25
3.2	XQuery Rules	26
4	Pronominalization Exercise	27
4.1	Examples	28
4.1.1	Accusative case	28
4.1.2	Dative case	29
4.2	Specific Exercise Architecture	30
4.3	Sentence Selection	31
4.4	Complement Selection and Analysis	33
4.4.1	Gender and Number Selection	34
4.5	Pronoun Case and Form Generation	35
4.6	Pronoun Positioning Rules	35
4.6.1	Rule 1: Simplest case of affirmative main clauses without verbal chains	36
4.6.2	Rule 2: Verbal chains	36
4.6.2.1	Clitic Positioning within verbal chains: Empirical Study	37
4.6.3	Rule 3: Clitic attraction by negation	37
4.6.4	Rule 4: Indefinite and negative subjects	38
4.6.5	Rule 5: Clitic-attracting adverbs	38
4.6.6	Rule 6: Subordinate clauses	39
4.7	Response Generation	39
4.7.1	Distractor Generation	39
4.8	Exercise Interface	40
4.8.1	Question Interface	40
4.8.2	Feedback Interface	40

5	Evaluation	43
5.1	Evaluation Setup	43
5.1.1	Expert Analysis	43
5.1.2	Expert Evaluation Measures	44
5.1.3	Crowd-sourced Testing	45
5.1.4	Questionnaire	45
5.2	Expert Analysis Results	47
5.3	Crowd-sourced Test Results	49
5.3.1	Native Speakers Results	49
5.3.1.1	NS Questionnaire Results	50
5.3.2	Non-Native Speakers Results	55
5.3.2.1	NNS Questionnaire Results	55
5.3.3	Questionnaire Comments	56
6	Conclusion and Future Work	61
6.1	Final Remarks	61
6.2	Future Work	62
I	Appendices	69
A	XQuery Rule Example	71
B	Clitic Positioning within verbal chains: Empirical Study	73
C	Questionnaire	77

List of Figures

2.1	REAP.PT architecture adapted from (Marques, 2011).	3
2.2	REAP.PT ‘Lexical Mahjong’ exercise.	6
2.3	REAP.PT ‘Choice of mood in subordinate clauses’ exercise.	7
2.4	REAP.PT ‘Nominal Determinants’ exercise.	8
2.5	REAP.PT ‘Nominal Determinants’ feedback system.	8
2.6	REAP.PT ‘Collective Names’ exercise.	9
2.7	REAP.PT syntactic exercises architecture.	10
2.8	<i>Ciberescola</i> web-page.	11
2.9	Pronominalization exercises from <i>Ciberescola</i>	20
2.10	Example exercise from <i>Diálogos 7</i> (Costa & Mendonça, 2011).	21
2.11	Example of incorrect use of pronouns (<i>achas-te</i> instead of <i>achaste</i>).	21
3.1	REAP.PT new syntactic exercises architecture.	24
4.1	Exercise question interface.	41
4.2	Exercise feedback interface.	42
4.3	Exercise feedback interface with tool-tip on mouse-hover.	42
5.1	Exercise evaluation website introduction.	46
5.2	Exercise evaluation website user form.	47
5.3	Distribution of incorrect answers by distractor type for NS.	51
5.4	Results for the statement “The system is easy to use” for NS.	52
5.5	Results for the statement “I understood the objective quickly” for NS.	52

5.6	Results for the statement “The exercises are too easy” for NS.	53
5.7	Results for the statement “The presented feedback is sufficient” for NS.	53
5.8	Results for the statement “The system is useful: I learned something by using it” for NS. .	54
5.9	Results for the statement “Global system appreciation” for NS.	54
5.10	Distribution of incorrect answers by distractor type for NNS.	56
5.11	Results for the statement “The system is easy to use” for NNS.	57
5.12	Results for the statement “I understood the objective quickly” for NNS.	57
5.13	Results for the statement “The exercises are too easy” for NNS.	58
5.14	Results for the statement “The presented feedback is sufficient” for NNS.	58
5.15	Results for the statement “The system is useful: I learned something by using it” for NNS.	59
5.16	Results for the statement “Global system appreciation” for NNS.	59

List of Tables

4.1	Pronominal case in Portuguese	27
5.1	Total number of generated exercises	43
5.2	Number of generated exercises for sentences with less than 20 words	44
5.3	Evaluation precision for each rule.	48
5.4	Incorrect exercises by error class.	49
5.5	Incorrect answers by rule for NS.	49
5.6	Number of exercises deemed erroneous by the NS users.	50
5.7	Incorrect answers by rule for NNS.	55
B.1	Clitic positioning counts on auxiliary verbs with linking prepositions.	74
B.2	Clitic positioning counts on auxiliary verbs without linking prepositions.	75

Acronyms

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

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

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

DOM Document Object Model is a cross-platform and language-independent interface dynamically access and update the content, structure and style of HTML, XHTML and XML documents

EHU Euskal Herriko Unibertsitatea, the University of the Basque Country is a public university and the main research institution in the Basque Country, in Northern Spain.

FAST Free Assessment of Structural Tests) is an automatic generation system for grammar tests

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

INESC-ID Lisboa 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

L²F Spoken Language Systems Laboratory is a research department at INESC-ID Lisboa

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

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

NNS Non-native speakers

NS Native speakers

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

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

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

STRING STatistical and Rule-based Natural lanGuage is an NLP processing chain for Portuguese developed a L2F

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

TESOL Teaching English to Speakers of Other Languages is an association whose mission is to advance professional expertise in English language teaching and learning for speakers of other languages worldwide, and provides teaching and learning standards

TOEFL Test of English as Foreign Language, an well-established and standardized multiple-choice test

VOA Voice of America is the official external broadcast institution of the United States federal government

XIP Xerox Incremental Parser is an on-the-fly rule compiler, with syntactic and semantic text parsing functionalities.

XML Extensible Markup Language is a specification defined by W3C, which allows to extensively create markup languages, with the main purpose of sharing structured data between different information systems

XPath XML Path Language is a query language for selecting nodes from an XML document

XQuery XML Query is a query and functional programming language that is designed to query collections of XML data

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

1 Introduction

In the last decades, an increased appearance of targeted and adapted products has been seen replacing mass-oriented and generic ones in many areas, including advertising, news and information, and, recently, even “Personalized Medicine”¹ is being researched and applied. Technology has changed how people use and treat information, making them to expect increasingly personalized and dynamic information systems, as opposed to the static and generic means of obtaining and processing information of the past.

In the education area, these trends also apply and have had a high impact in the learning process, where attention and motivation are of utmost importance, and teaching materials must be appealing to the students.

It is in this context that the Computer Assisted Language Learning (CALL) research area has appeared, with the aim of developing tutoring tools adapted to the students’ expectations and their specific needs, and thus improving the learning process.

The REAP (REAders-specific Practice) project² is one of such systems, developed at CMU³ by the LTI⁴ for the teaching of the English language. It aims at teaching vocabulary and practice reading skills (lexical practice), using dynamic games and exercises, adapted to each student learning level and interests, helping teachers to target and accompany each student individually. It uses real documents extracted from the web, providing recent, varied, and thus more motivating reading material.

Automatic exercise generation, one of the most important and differentiating features of REAP, is made possible by the application of computational linguistics, which is one of the characteristics of the specialized CALL systems in the emerging interdisciplinary field of Intelligent Computer-Assisted Language Learning (ICALL)⁵.

The REAP.PT⁶ project aims to bring the REAP learning strategies to the Portuguese language. The lexical learning component, analogue to the original REAP system, is comprised of the text reading

¹http://en.wikipedia.org/wiki/Personalized_medicine (last visited in October 2012)

²<http://reap.cs.cmu.edu> (last visited in October 2012)

³Carnegie Mellon University - <http://www.cmu.edu> (last visited in October 2012)

⁴Language Technologies Institute - <http://www.lti.cs.cmu.edu> (last visited in October 2012)

⁵<http://purl.org/calico/icall> (last visited in October 2012)

⁶<http://call.l2f.inesc-id.pt/reap.public> (last visited in October 2012)

and question generation phases, and was developed in Marujo (Marujo, 2009) and Correia (Correia, 2010). More recently, a listening comprehension module was also developed (Pellegrini et al., 2011). The system was then extended to include syntax learning as well, starting in Marques (Marques, 2011), and continued in the present work.

1.1 *Goals*

The goal of the present work is to continue the development of the syntactic module of the REAP.PT tutoring system, through the development of additional exercises. These exercises should exhibit the same features that make the tutoring tool compelling to both students and teachers. Namely, they should be automatically generated and use real texts as source.

In this context, a new module of exercises was developed in this project, focusing on the pronominalization of syntactic constituents. This exercise is often presented in grammar drills in Portuguese textbooks, and also constitutes a challenging aspect for language learners. The proposed exercise is explained in more detail in chapter 4.

1.2 *Document Structure*

The present thesis consists of 6 chapters, and it is structured as follows:

- Chapter 2 starts by introducing the REAP.PT system, describing its architecture and currently implemented exercises. It then presents the state of the art for manually and automatically generated language learning exercises.
- The general exercise generation architecture of this work is presented in Chapter 3.
- The general architecture presented in Chapter 3 provides the basis for the creation of the pronominalization exercise described in Chapter 4. The exercise is explained in detail with examples, and the several generation steps and rules are described, ending with the interface modules.
- Chapter 5 is about the evaluation setup and results, involving an expert analysis and crowd-sourced testing of the exercise.
- Finally, Chapter 6 presents the conclusions of this thesis and suggests future work.

2

State of the Art

In this chapter, the state of the art review is presented, starting with the current state of the REAP.PT project. Other Portuguese CALL systems with similar characteristics to REAP are described, although none of them has automatic question generation. Automatic generation systems for other languages are also described. Finally, a brief overview of PFL textbooks is made to present the most common variations of the exercises here proposed. This also encompasses several Portuguese CALL systems.

2.1 REAP.PT

2.1.1 REAP.PT Architecture

This section describes the current architecture of the REAP.PT system (Correia, 2010; Marques, 2011), focusing on vocabulary exercises (see Figure 2.1).

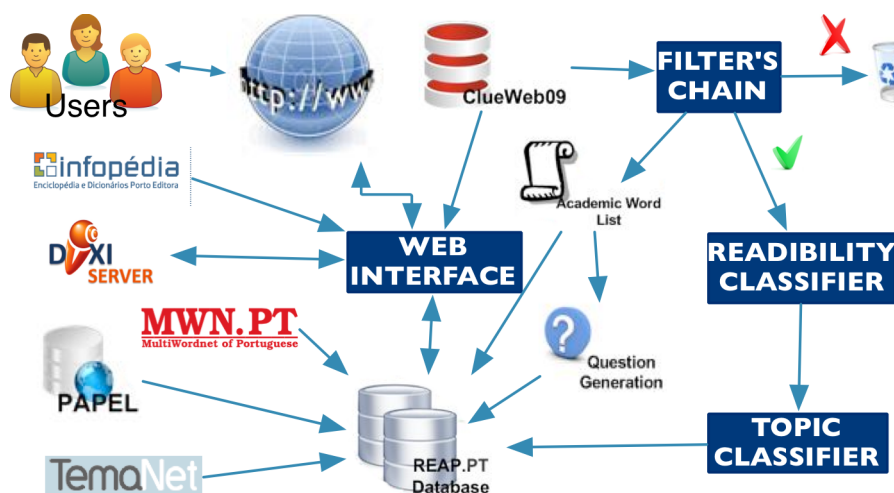


Figure 2.1: REAP.PT architecture adapted from (Marques, 2011).

The Web Interface component is responsible for:

- the user interaction with the system;
- information exchange between the database and the listening comprehension module; and

- getting dictionary definitions from Infopédia¹, the on-line dictionary of Porto Editora.

The listening comprehension module, represented in Figure 2.1 by the DIXI Server logo (Paulo et al., 2008), provides text-to-speech audio playback of text presented to the user, so that the students can also train their understanding of the spoken language.

The database module is divided in two relational databases. The first, specific to REAP.PT, contains the system state such as user information (topics of interest, proficiency level, readings and assessment history etc.), text information, focus words and related questions and distractors (these are generated prior to assessment). The second database stores the lexical resources from PAPEL² (lexical ontology with word relations), MWN.PT³ (wordnet with synonym sets) and TemaNet⁴ (wordnet and semantic domains).

The document corpus in use for vocabulary questions is a subset of the ClueWeb09, which consists of over 1 billion web pages in 10 languages, compiled by the Language Technologies Institute at CMU in 2009. In REAP.PT, only the Portuguese section is used, containing 37,578,858 web pages. To build ClueWeb09, the Nutch Crawler (Moreira et al., 2007) was used to extract the web pages.

A filter chain is used to select a subset of the corpus that fits within certain practical and pedagogical constraints (Marujo, 2009). The filters are presented in order of execution:

- a filter to eliminate short texts (with less than 300 words), and which also stores all the accepted texts' word count in the database;
- a filter to eliminate texts that include profanity words;
- a filter to eliminate texts with no valid sentences (lists of words); and
- a filter to eliminate texts that do not have at least three focus words present in the Portuguese Academic Word List (P-AWL) (Baptista et al., 2010)), that also identifies focus words in the accepted texts.

The topic and readability classifiers run on the output of the filter chain and classify the texts according to topic and reading level (Marujo, 2009).

The question generation module is responsible for the generation of vocabulary exercises given to the students after each text reading. The existing exercises include definition questions, synonym questions, hyperonym/hyponym questions, and cloze questions about the text. All these exercises involve

¹<http://www.infopedia.pt> (last visited in October 2012)

²<http://www.linguateca.pt/PAPEL> (last visited in October 2012)

³<http://mwnpt.di.fc.ul.pt> (last visited in October 2012)

⁴<http://www.instituto-camoes.pt/temanet> (last visited in October 2012)

multiple-choice questions and thus the generation of appropriate distractors (Marujo, 2009; Correia, 2010).

2.1.2 REAP.PT Exercises

The work on the question generation module started in Correia (Correia, 2010) with vocabulary questions, with a focus on *cloze* questions, also known as *fill-in-the-blank* questions, and the study of the distractors, the wrong alternatives in the multiple-choice cloze question, that distract the student from the right answer to provide him a challenge.

The questions are generated according to two aspects of the student model: level of proficiency and topics of interest. The level represents the student's Portuguese language proficiency approximated by the vocabulary knowledge he/she displays (REAP's basic feature in vocabulary learning). It is estimated on the first time the student uses the system, and it evolves according to the student's activities, such as reading sessions, dictionary lookups and assessments. The interests of each student are recorded using a survey, according to a set of categories. The system then prioritizes the set of texts presented to the student for reading and exercising, according to his/her preferences.

The vocabulary learning is focused on a specific set of words that the student should learn, adapted to the student's level. These are called *focus words*, and constitute a sub-set of the Portuguese Academic Word List (P-AWL) (Baptista et al., 2010). The P-AWL (Baptista et al., 2010) is defined "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."


There are two types of definition exercises. The first, developed in (Correia, 2010), are multiple-choice questions generated from dictionary definitions and the distractors randomly chosen from the remaining P-AWL words⁵.

The second exercise, developed in (Marques, 2011), is a ludic game based on Mahjong Solitaire called 'Lexical Mahjong', in which the student has to establish a correspondence between the lemma and the definition of a word. It uses a filter chain for the selection of the word-definition pairs. These filters remove definitions containing words cognate of the target word; eliminate longer definitions, above a predefined length; and clear certain typographic elements that might hinder comprehension. Finally, a set of classifiers are used to determine difficulty level and whether the definition belongs to a specific scientific/technical domain. Being a ludic exercise, a scoring mechanism was added, for the first time in REAP.PT. The student is given a set of points according to the resolution of the exercise, so

⁵More recently, (Correia et al., 2012) used ML techniques to improve the quality of the cloze questions stems (target sentences)

both the student and teacher can have feedback of the student's performance, and add motivation for solving the exercise quickly and without making mistakes (hesitations and elapsed time penalize the score). An example of this exercise can be seen in Figure 2.2.

Mahjong Lexical



Pontos 80

Palavras			Definições	
enorme	ilegalidade		clarificação de dúvida(s)	forma de organização de informação em linhas e colunas
tabela	esclarecimento		acto ou situação contrários à lei	que é muito grande

Figure 2.2: REAP.PT 'Lexical Mahjong' exercise.

The synonym questions (Correia, 2010) are generated from the most common relations of the focus words in the resources, and distractors are selected from words with the same POS and level of classification as the word being tested.

For the vocabulary *cloze* questions, sentences of the text read by the student are presented (stem), in which one word is removed leaving a blank space, while the student has to choose the right word to fill the blank. The word removed from the stem is one of the focus words that the student has to learn, and for that purpose all the inflections of the P-AWL words are used. The distractors for cloze questions are generated using two approaches, the graphemic distractors (P-AWL words with the same POS and lowest Levenshtein Distance) and phonetic distractors (common spelling errors of the word).

The current syntactic exercises in REAP.PT (Marques, 2011) are the 'Choice of mood in subordinate clauses' exercise and the 'Nominal Determinants' exercise.

The ‘Choice of mood in subordinate clauses’ exercise aims to teach the syntactic restrictions imposed by the subordinative conjunctions on the mode of the subordinate clause they introduced. The rule-based parser XIP-PT (N. J. Mamede et al., 2012), based on XIP (Aït-Mokhtar et al., 2002), creates sub-clause chunks that link the conjunctions and conjunctive locutions (previously recorded in the system lexicon) to the first verb of the subordinate clause. Distractors are then generated using the L²F VerbForms⁶ word form generator for verbs, and a set of rule-based restrictions are applied to reduce ambiguity. An example of this exercise can be seen in Figure 2.3.

Figure 2.3: REAP.PT ‘Choice of mood in subordinate clauses’ exercise.

The ‘Nominal Determinants’ exercise aims to teach distributional constraints between a determinative noun and the noun it determines (e.g. *copo de leite*), and at the same time the relationship between collective names and common names (when collective names function as determinative nouns on common nouns, e.g. *mata de cedros*). Quantifying dependencies are detected in the sentences taken from the corpus, holding between the nominal or prepositional phrase containing the determinative noun and the subsequent prepositional phrase containing the determined noun. The determinative noun (target word) is then removed from the sentence, and distractors are generated from a list of determinative and collective names previously added to the lexicon. Semantic features of the nouns are used to avoid generating correct answers that share the same semantic category with the target word. Figuratively associated categories are also ignored (such as Human and Animal, e.g. *alcateia de políticos*), to avoid ironic relationships. Generic determinative nouns (e.g. *conjunto*, *grupo*) are also discarded. A feedback system teaches the student the missed definitions, giving examples and images illustrative of the determinative nouns. An example of this exercise can be seen in Figure 2.4, along with the feedback system in Figure 2.5.

To simplify the learning of collective names separately from other determinative nouns, a new

⁶<https://www.l2f.inesc-id.pt/wiki/index.php/VerbForms> (last visited in October 2012)

⁷This exercise exhibits one problem with the distractors, the lack of concordance in gender with the right answer, making these distractors less convincing than desired.

Selecione a palavra que melhor complete a frase.

O do carro pequeno passa-lhe lhe um chorudo _____ de notas.

☐ imensidade
☐ ânfora
☐ maço
☐ canastra

Seguinte

Figure 2.4: REAP.PT ‘Nominal Determinants’ exercise.⁷

A sua resposta está errada!!

O do carro pequeno passa-lhe lhe um chorudo **ânfora** de notas.

Um exemplo de utilização da palavra ânfora:

/ Toma uma **ânfora** de vinho, senta-te te ao luar e bebe, / lembrando-te te que, talvez amanhã, a lua te procurará em vão.

Imagens ilustrativas da palavra ânfora:

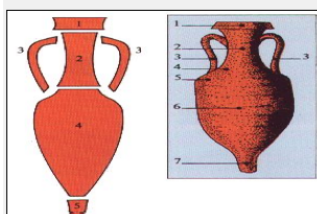


Figure 2.5: REAP.PT ‘Nominal Determinants’ feedback system.

exercise called ‘Collective Names’ was created on the REAP.PT interface, selecting questions from the ‘Nominal Determinants’ exercise with only collective names as correct answers. An example of this exercise can be seen in Figure 2.6.

2.1.3 REAP.PT Syntactic Exercises Architecture

The general architecture of the syntactic exercise generation can be seen in Figure 2.7.

The exercises are generated from the CETEMPúblico⁸ corpus (Santos & Rocha, 2001) processed by the STRING⁹ (N. J. Mamede et al., 2012) processing chain of L²F, instead of using the texts presented to

⁸<http://www.linguateca.pt/CETEMPUBLICO> (last visited in October 2012).

⁹<https://string.l2f.inesc-id.pt> (last visited in October 2012).

Selecione a palavra que melhor complete a frase.

Divulgadas pelo PÚBLICO e proferidas durante as investigações do massacre do Meia Culpa, aquelas declarações levaram o IGAI a deslocar a Amarante uma _____ de investigadores.

☐ bosque
☐ turma
☐ comunidade
☒ equipa

[Seguinte](#)

Figure 2.6: REAP.PT ‘Collective Names’ exercise.

the student from the ClueWeb09 corpus and used in the remaining modules of REAP.PT. This is done in order to provide more text and variety to the student, and for the higher quality of the text which not only helps the generation of syntactic exercises, but also are more adequate than ClueWeb09 for the pedagogical purposes of REAP.PT.

The result from the syntactic analysis of the corpus (output of the XIP-PT parser) consists of XML files containing the syntactic tree of each sentence and the syntactic dependencies between the sentences’ nodes.

In the sentence selection phase, the XIP output is processed, and the syntactic features are analyzed in order to select the stems that are to be used to generate the questions. This phase is performed using the Hadoop¹⁰ Map-Reduce framework for distributed processing, in order to reduce the processing time. In each map operation one sentence is processed, using the DOM (Document Object Model), which represents the XML in a tree structure that is then traversed recursively, using flags when a relevant dependency is found. Despite needing to analyze only one dependency to generate the existing exercises, the code becomes complex and thus difficult to maintain. Each exercise has it’s own sentence selection program, with no shared code between exercises. The selected sentences are output in plain text. Since the exercises use cloze questions, a blank space replaces the selected correct word. Some syntactic information about the chosen word is also appended to the sentences, needed for the distractor generation phase. Separate programs for each exercise introduce the selected sentences into the database, in separate tables.

In the “Choice of mood in subordinate clause” exercise, the distractors are generated using a specific program, using the correct-answer words extracted from the sentences, as described in section 2.1.2. The distractors are then inserted into the database. In the “Nominal Determinants” exercise, on the other

¹⁰<http://hadoop.apache.org> (last visited in October 2012)

hand, the distractors are generated on-the-fly in the web interface module during presentation. This is done because the distractors are chosen from a list of words, so there is no need to generate them beforehand.

In this section, a general overview of the REAP.PT current development was presented, focusing on the main exercises that have already been implemented. In the next section, other CALL systems currently available for Portuguese are briefly described.

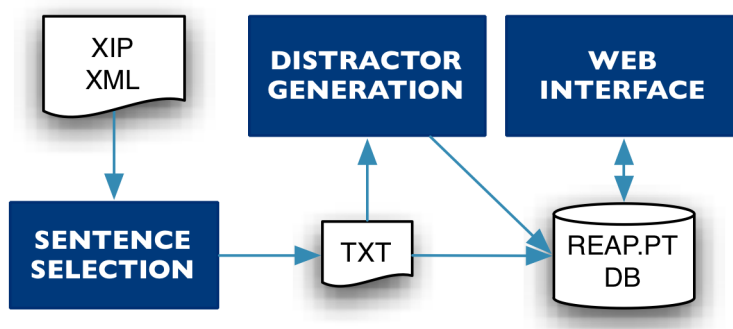


Figure 2.7: REAP.PT syntactic exercises architecture.

2.2 Portuguese CALL Systems

We present two CALL systems for the Portuguese language, which are available online, and, like REAP.PT, have web interfaces. The first, Ciberescola, also has some pronominalization exercises, shown in section 2.4, Figure 2.9.

2.2.1 Ciberescola

The ‘Ciberescola da Língua Portuguesa’¹¹ (Cyberschool of the Portuguese Language) is a platform of interactive resources and online courses for Portuguese teaching, opened since September 2011 to the general public. Portuguese native students (levels 5º to 12º) and Portuguese-as-second-language students (levels A1 to C2) have at their disposal interactive exercises (about 1000 at the moment) ranging several language proficiency areas (reading, oral comprehension, grammatical, writing and vocabulary), and organized by student level and difficulty level (easy, normal, hard).

All exercises are original, and were “manually” produced by teachers and researchers in linguistics, literature and language teaching. Therefore, they are not automatically generated nor are they adapted to the students topics of interest as with REAP. Instead, exercises focus on the addressed competences.

¹¹<http://www.ciberescola.com> (last visited in October 2012).

ciberescola)
da língua portuguesa

(início) (sobre nós) (guiões) (cibercursos) (dicionários) (FAQ) (equipa) (contactos) (sair)

Tiago Freitas
nível C2

a minha conta
mudar de nível
sair

Exercícios

ciberescola)
uma iniciativa do
CIBERDÚVIDAS
DALÍNGUA PORTUGUESA

MINISTÉRIO DA EDUCAÇÃO E CIÊNCIA

Observatório
da Língua Portuguesa

SHARE

Voz passiva e voz ativa

1. Passa as seguintes frases para a voz passiva:

Pontuação: 0 1 respostas certas 39 respostas erradas

Nível C2

Área: Gramática
Grau: normal

1. O grupo C realizou o trabalho mais completo. **O trabalho mais completo foi realizado pelo grupo C.**

2. Durante o assalto, os bandidos mataram um polícia. <não preenchido>

3. O Filipe escreveu uma carta à prima. <não preenchido>

4. A D. Helena limpou estas duas janelas. <não preenchido>

5. No próximo mês, dois cientistas portugueses iniciarão uma importante viagem de pesquisa. <não preenchido>

6. A polícia soltou três presos ontem à noite. <não preenchido>

7. O Paulo descobriu um inseto raro. <não preenchido>

8. O tio pagou os bilhetes do cinema. <não preenchido>

9. O Serafim comprou um fato novo. <não preenchido>

10. O José e a Paula viram um grupo de turistas no museu. <não preenchido>

11. Ontem, a polícia prendeu um perigoso malfeitor. <não preenchido>

Figure 2.8: Ciberescola web-page.

Ciberescola has several types of exercises, including true/false questions, cloze questions with distractors, fill in the blanks, open answers (for example in the transformation of active/passive sentences), and correspondence. Each exercise has a clue (*dica*) in the beginning, explaining how the exercise works and providing an example with the correct answer. Images relevant to the exercise also appear alongside the text or the answers, increasing the appealing effect to the student and helping him/her to understand the content. There is a suggested order to resolve the exercises, so the student can improve his competences gradually, with the interdependence of subjects in mind. It is also possible for the student to see the exercises that s/he has already tried in the past, with the classifications and marks telling which exercises should be tried again because of insufficient results. The corrections are automatic, and in the case of open answers, it is supposed that all possible correct answers were previously manually stored in the system¹².

When the answer is wrong, there is no further analysis to indicate where the error lies.

¹²Information about this process was not available on the site.

2.2.2 Aprender Português

'Aprender Português'¹³ is an area of the '*Centro Virtual Camões*' (Camões Virtual Center), of *Instituto Camões* (Camões Institute), a public institute with the mission of internationally promoting the Portuguese language and culture.

It currently features several resources for Portuguese learning (speaking, listening and reading skills), and has announced that, in the future, it will also present exercises for training writing skills.

These resources include a didactic games section, listening comprehension, reading comprehension and conversation audio guides. All exercises are organized in three difficulty levels, and most require the Flash browser plug-in to function. All the exercises are static and they are not automatically generated.

The games section has several ludic exercises:

Lusophone Game A multiple-choice ludic exercise with questions about the lusophone (Portuguese-speaking) countries, with several themes (history, culture, etc.).

Hangman Game Traditional game to test word memory.

Glory Game Traditional board game with multiple-choice language questions.

Lexical exercises Several types of exercises to learn vocabulary, expressions, synonyms, etc. They include association exercises, crosswords, and multiple-choice exercises.

Grammar exercises Several association and fill-in-the-blank (multiple-choice) exercises on basic grammar.

The reading comprehension section has several texts and books, with multiple-choice and fill-in-the-blank exercises about their content.

2.3 ICALL Systems

There are not many ICALL systems that include automatic generation of exercises, and even less for syntactic exercises, as it can be seen in previous state of the art survey on this topic (Marques, 2011). In this section, a summary of the last state of the art survey on this subject (Marques, 2011) is presented, augmented with the latest information available on those projects.

Two additional and important systems are reviewed: FAST, that specializes in grammar exercises, and ArikIturri, a general multilingual system.

¹³<http://cvc.instituto-camoes.pt/aprender-portugues.html> (last visited in October 2012).

2.3.1 TAGARELA

TAGARELA (Teaching Aid for Grammatical Awareness, Recognition and Enhancement of Linguistic Abilities)¹⁴ (Amaral & Meurers, 2011) is an ICALL system for the Portuguese language, developed by the ICALL Research Group¹⁵ at Ohio State University, and further developed at Tübingen University's Department of Linguistics¹⁶.

The student using the system can practice listening, reading and writing skills, with feedback on spelling, morphological errors (non-words, spacing, capitalization, punctuation), syntactic errors (nominal and verbal agreement), and semantic errors (missing or extra concepts, word choice).

TAGARELA's exercises include listening and reading comprehension, description of pictures and text, vocabulary practice (in the form of fill-in-the-gap exercises), and re-phrasing.

Other than REAP.PT, TAGARELA is the only ICALL system for Portuguese found in this review. Although it does not provide automatically generated syntactic exercises, this system provides feedback for syntactic errors in written text.

2.3.2 Working With English Real-Texts

The WERTI (Meurers et al., 2010) system processes real texts in order to generate several syntactic exercises for the English language. Like TAGARELA, it was developed by the ICALL Research Group at Ohio State University, and also further developed at Tübingen University's Department of Linguistics.

It uses a rule-based NLP chain, and has a web interface¹⁷. Contrary to REAP.PT, it does not use a corpus of previously filtered web pages, opting instead for processing any web page the user selects. Recently, a Firefox plug-in was developed, leaving only the NLP up to the server. This was done to increase compatibility with web pages using dynamically generated contents and special session handling.

Besides the original English language, Spanish and German are now working in a beta phase.

It has two types of exercises, the so-called *Click* activities and *Practice* activities. In the *Click* activities, the students have to identify patterns (such as grammatical categories) through clicking and automatic color feedback. The *Practice* activities consist of fill-in-the-blank exercises with no distractors (only the form in the original text is accepted). It also includes word order rearrangement exercises.

The implemented exercises are comprised of lexical category identification (with a focus on prepositions and determiners), gerund/infinitive application, conditionals and phrasal verbs.

¹⁴<http://purl.org/icall/tagarela> (last visited in October 2012)

¹⁵<http://www.ling.ohio-state.edu/icall> (last visited in October 2012)

¹⁶<http://www.sfs.uni-tuebingen.de> (last visited in October 2012)

¹⁷<http://purl.org/icall/werti> (last visited in October 2012)

This system does not have distractor generation, it does not include multiple-choice questions, and the fill-in-the-blank exercises only accept one correct answer (out of several possible answers in some cases, like different prepositions, a problem identified in the paper). Since it does not manipulate the original texts to generate exercises, it currently has a more limited scope in exercise generation than REAP.T, focusing instead on the fact that it can be easily applied to any web page in a short period of time without interrupting the web browsing experience.

2.3.3 The Alpheios Project

Alpheios Reading Tools¹⁸ is an open-source project by The Alpheios Project non-profit organization.

This project's goals is to develop a language learning software that can be adapted to a student's specific goals and needs. The supported languages are Latin, Ancient Greek and Arabic, with Chinese and Spanish still in development.

For any text that the user uploads or from a web page, this application provides: word definitions, word morphology, inflection tables, and a personal word list manager (so the student knows which words were learned before). Students can also create *Personal Sentence Diagrams*, which are syntactic trees that can be edited and annotated by the student. Inflection and vocabulary frequency analysis is also available for any text.

The system presents additional features for *Enhanced Texts*, a collection of pre-processed texts using an NLP chain. These features are *Aligned Translations*, *Sentence Diagrams*, which presents syntactic trees, and *Quizzes*, a multiple-choice exercise in which the student has to classify each word according to its part-of-speech (POS), the correct translation to a target language and its form (number, gender and case).

2.3.4 FAST

FAST (Free Assessment of Structural Tests) is an "automatic generation system for grammar tests" (Chen et al., 2006).

This system generates grammar exercises for the English language, using a method that involves representing the questions' characteristics as structural patterns (surface patterns), acquiring authentic sentences on the Web, and applying those patterns in order to transform sentences into exercises questions. Sentences are converted into two types of questions: traditional multiple-choice cloze questions, and error detection questions, in which several slots (groups of words) on the sentence are marked; and the student has to identify the incorrect slot.

¹⁸<http://alpheios.net> (last visited in October 2012)

The surface patterns are made of POS tags that can, for example, specify certain specific verb tenses. They can be structural patterns for the question generation, or distractor patterns. Another area where surface patterns are used is in semantic question-answering (Q\A) multiple-choice tests, where lexico-syntactic patterns can be used that relate questions with answers (Mendes et al., 2011).

One example of a pattern, taken from (Chen et al., 2006) is the following:

**X/INFINITIVE * PP.*
 →
** _____ * PP.*
 (A) X/INFINITIVE
 (B) X/to VBG
 (C) X/VBG
 (D) X/VB

This pattern allows the generation of this kind of question:

Representative democracy seemed _____ simultaneously during the eighteenth and nineteenth centuries in Britain, Europe, and the United States.
 (A) to evolve
 (B) to evolving
 (C) evolving
 (D) evolve

The distractors for the infinitive verb are thus generated using other forms of the same verb.

Since distractors are “usually some words in the grammatical pattern with some modification” (changing part of speech, adding, or deletion of words), several symbols can be used to designate specific words: \$0 for the target (key) word, and \$9 and \$1 for the word preceding and following the target word respectively.

The patterns used in the evaluation of the system are made using test patterns adapted from TOEFL (Test of English as Foreign Language), an well-established and standardized multiple-choice test.

The concept of question ‘formation strategies’ is also used, as a way to describe the processes of generating different types of question: traditional multiple-choice and error correction questions.

An evaluation was performed, in which 69 test patterns were constructed by adapting a number of grammatical rules from TOEFL, covering nine grammatical categories. Sentences from Wikipedia and VOA (Voice of America broadcast news) were matched against the test patterns, and transformed

into multiple-choice and error detection questions. "A large amount of verb-related grammar questions were blindly evaluated by seven professor/students from the TESOL program. From a total of 1,359 multiple-choice questions, 77% were regarded as 'worthy' (i.e., can be direct use or only needed minor revision) while 80% among 1,908 error detection tasks were deemed to be 'worthy'." (Chen et al., 2006).

2.3.5 Arikiturri

Arikiturri (Aldabe et al., 2007; Aldabe, 2011) is a multilingual automatic question generation system, developed at the IXA research group¹⁹, at the University of the Basque Country (EHU), in the context of a PhD thesis²⁰. It is currently implemented for Basque language learning, English language learning and science domains. It can generate several types of questions: error correction, fill-in-the-blank, word formation, multiple-choice and short answer questions.

Taking the abstraction concepts present in the FAST system one step further, it uses a question model to represent the exercises (as well as the information relating to their generation process) in a general and flexible way. "It is a general model because of its independence from the language of the questions as well as from the NLP tools used for their generation. [...] [It] allows different types of questions to be represented and, in addition, different types of questions can be specified into the same exercise. Finally, because the model has been developed using XML, the importation and exportation processes [into independent applications] are easy tasks. [...] [The] model is also flexible due to different reasons. First of all, [...] new types, such as word order and transformation, could be also represented by this model. Besides, it also offers the possibility of changing the order of the chunks in a sentence." (Aldabe et al., 2007).

In this question model, an exercise is a set of questions. The question is composed of the **topic**, the **answer focus** and the **context**. The answer foci are the chunks of the sentence where the topic appears. The rest of the chunks are put into the context. The answer focus consists of a **head** and a **notHead**. Only the head contains the necessary information of the chunk to treat the topic. This representation allows to change the order of the chunks of the sentence. The **change** attribute delimits which chunks can undergo order changes. The head is divided into the answer, a list of distractors and a list of **headComponents**. The answer is the minimum list of words where the topic appears, the topic info and the analysis related to it. Distractors are always linked to an answer focus. The corresponding linguistic analysis and the heuristics used for creating them are also stored. Finally, the **headComponent** collects the specific information related to the question type. There is also a **rule** attribute in order to explain how each headComponent is created. (Aldabe et al., 2007)

¹⁹<http://ixa.si.ehu.es/Ixa> (last visited in October 2012)

²⁰<http://ixa2.si.ehu.es/ialdabe/phd.html> (last visited in October 2012)

A web-based post-editing environment was also developed, in order to evaluate, manually, the generated questions. Post-editors can accept, discard or modify questions. Both the source sentence and the distractors can be modified, and the reasons for the modifications can be added so the system can be improved. The information used in the generation process is available to the post-editor, since it is represented in the question model.

The system is highly modular, which contributes to ease the process of adding a new type of exercise, feature or heuristic, and to its multilingualism. The use of object-oriented programming contributes to this modular and re-usable design.

"The process of generating test items can be summarised as follows: based on the parameters' specifications, the sentence retriever module selects candidate sentences from the source corpus which has been designated as the source. In the first step, it selects the sentences where the specified topic appears. Then, the candidate selector module, based on the defined criterion, selects the candidate sentences. Once the sentences are selected, the answer focus identifier marks out some of the phrases as focal points for the answers depending on the information contained within them. Then, the item generator creates the questions in accordance with the specified exercise type. This is why, this module contains the distractor generator sub-module. As the entire process is automatic, it is probable that some of the questions will be ill-formed. [...] For this reason, we included the ill-formed question rejecter module in the architecture." (Aldabe, 2011)

The system obtains intermediate results between the modules, making it possible to use the same test data to generate another type of exercise (for example, from multiple-choice to fill-in-the-blanks). The system would start from the item generator module instead of the sentence retriever module, due to these intermediate results.

2.4 *Current Syntactic Exercises on Pronominalization*

Here we present examples of the current exercises which were found in Portuguese textbooks and on-line resources, focusing on variations of the exercises proposed in this dissertation.

There are several pronominalization exercises in textbooks and on-line resources, shown here in a tentative order of difficulty:

1. Given three forms of pronouns (*lo(s)/la(s),no(s)/na(s),lhe(s)*), choose the right one to replace the signalled constituent. This is the easier type of exercise, since in this case, the student does not have to remember the pronouns nor its correct position on the sentence. He/she only has to select the correct form.

Example from Português 10 (VVAA, 2010):

Substitui as expressões a negrito pelos pronomes adequados:

- *Fizeste o trabalho sobre o texto da Lídia Jorge?*

- *Eu fiz **o trabalho**, aliás faço sempre **os trabalhos**. Se não fizesse **os trabalhos**²¹, não teria tão boa nota.*

*Queres ver **o meu trabalho**?*

- *Sim, queria ler **o trabalho**.*

22

2. Correct and incorrect sentences, that must be classified according to clitic placement (position on the sentence). The student doesn't have to remember the pronouns, and only has to identify their correct placement in the sentence. This type of exercise can be seen in the examples of the goals section 4.1. In those examples, only one correct sentence is shown among the distractors. One variation is showing several correct and incorrect sentences to be classified.

3. Cloze questions, sentences with a stem example, with a blank space for an expression that has been deleted, and in which the student has to insert a correct answer, choosing from a set of alternatives (multiple-choice exercises) or fill in the correct pronoun (fill-in-the-blanks exercises). In this case, the constituent that is to be replaced by a pronoun is signalled.

The multiple-choice cloze questions are easier to generate automatically, since they can be produced by modifying the original sentence in the text. We present two examples in Figure 2.9 (*a* and *b*). Since these examples are "manually" produced, they have context sentences followed by sentences in which the constituent is to be pronominalized. This contextualization is harder to generate automatically, because it involves generating entirely new sentences, instead of just manipulating the original sentences in a text.

4. Given a small text with signalled pronouns, rewrite the text replacing the pronouns with their corresponding antecedents. We present an example in Figure 2.10. This exercise is much more difficult to generate and assess, given that it involves the production of text. The identification of the antecedents in itself is a difficult problem that involves anaphora resolution, something that has not yet been addressed with sufficiently good results. In a recent system built for Portuguese (Nobre, 2011) only an f-measure of 33.5% was achieved. This is still an insufficient result for the purposes of this project since in ICALL it is imperative to minimize the number of errors that could be presented to the students, for it would compromise the learning process. The automatic assessment of the answer would also be very difficult, because it is not unusual for a pronoun to have more than one candidate antecedent in previous sentences, therefore anaphora resolution, even considering only pronominal anaphora, can be considered to be an open issue

²²In this case, both negation and subordination imply the fronting of the pronoun, so it is also a matter of positioning of the pronoun.

for Portuguese, and therefore that would require such techniques were discarded from the current project.

5. Given a declarative affirmative sentence with clitics, transform it to the corresponding negative sentence.

Example from *Português 10* (VVAA, 2010):

Atenta na frase:

A Andorinha ofendeu-o e ele vai agredi-la. O Gato Malhado disse-me que depois fá-la-ia pedir-lhe desculpa.

Experimenta agora pôr as mesmas orações na negativa. Que conclusões relativamente à colocação do pronome?

Notice the alternative positioning of the clitic in the first sentence, yielding two correct solutions: *e ele não a vai agredir / e ele não vai agredi-la*. This exercise is unclear whether all clauses are to be modified by negation or only the verbs of the main clauses. Notice, also, that there can be an alternative positioning of the mesoclitic *fá-la-ia*, since proclisis would also be accepted, as the pronoun can be attracted both for the subordinate context and the adverb: *que depois não a faria pedir-lhe desculpas*. Finally, if all verbs become modified by negation, that would render the sentence unacceptable (even incomprehensible).

This exercise aims to teach a specific pronoun positioning restriction imposed on the sentence by negative adverbs. However, sentence transformation exercises are difficult to generate automatically, and lead to formulation problems as seen here, where there can be more than one possible solution, making it very difficult to evaluate. Besides, the transformations are complex since the pronoun positioning rules involve several linguistic factors (presented in detail on section 4.6).

2.4.0.1 Common Student Errors

According to colloquial evidence reported by teachers, among the most common errors are the incorrect placement of the pronoun and the lack of use of the hyphen between the verbal ending and the clitic. Related to this last error type, it is also a common error the use of the past perfect tense of the verbs, where the last syllable is confused with an atonic pronoun, as can be seen in Figure 2.11 (e.g. *achaste/achas-te*). A similar error involves the use of imperfect subjunctive '*achasse/acha-se*'.

This mistakes can only be corrected by the use of different types of exercises, and appropriate distractors could be created from the target sentence to exercise them. For example, distractors could be made where the clitics appear adjacent to the verb missing the hyphen, and in incorrect positions.

Futuro + pronomes

Preenche os espaços com a forma correta do verbo no futuro, combinada com o pronome que substitui a expressão sublinhada.

DICA

Repara:

*Eu visitarei **a minha irmã** no Natal. – Eu visita-la-ei no Natal.*

*Tu comprarás **o casaco** quando vierem os saldos. – Tu compra-lo-ás quando vierem os saldos.*

*Nós vamos telefonar **ao João** quando ele chegar. – Nós telefonar-lhe-emos quando ele chegar.*

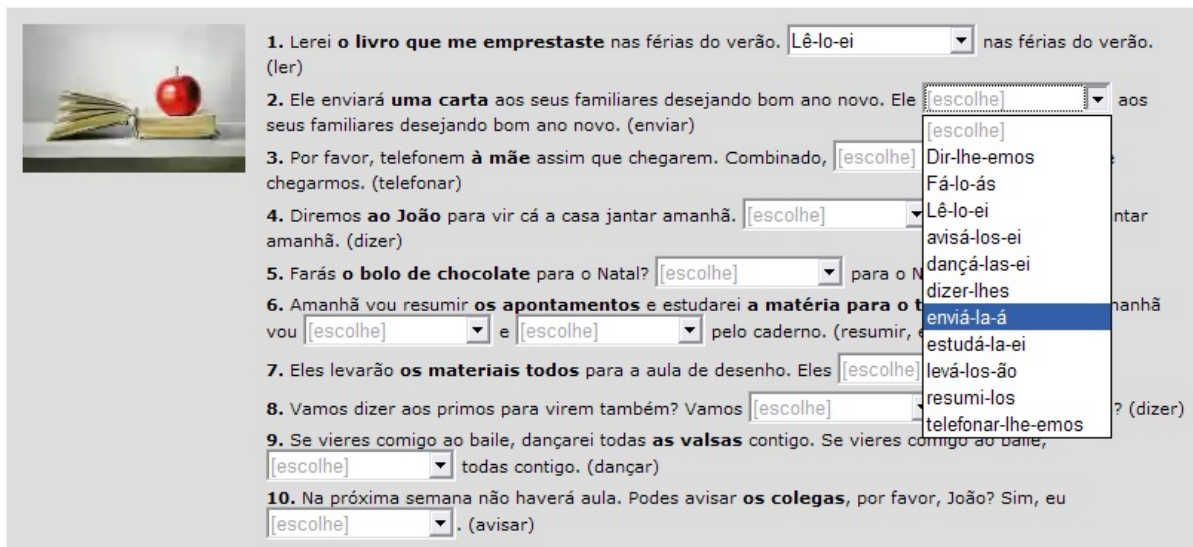
Muitas vezes usa-se o verbo «ir» no presente do indicativo + outro verbo no infinitivo para transmitir a ideia de uma ação futura. Nestes casos, os pronomes aparecem sempre associados à forma verbal, no final.

Nível C1

G

Área: Gramática

Dificuldade: difícil



1. Lerei o **livro que me emprestaste** nas férias do verão. Lê-lo-ei nas férias do verão. (ler)

2. Ele enviará **uma carta** aos seus familiares desejando bom ano novo. Ele [escolhe] aos seus familiares desejando bom ano novo. (enviar)

3. Por favor, telefonem **à mãe** assim que chegarem. Combinado, [escolhe] chegarmos. (telefonar)

4. Diremos **ao João** para vir cá a casa jantar amanhã. [escolhe] jantar amanhã. (dizer)

5. Farás **o bolo de chocolate** para o Natal? [escolhe] para o Natal. (fazer)

6. Amanhã vou resumir **os apontamentos** e estudarei **a matéria para o teste**. Vou [escolhe] e [escolhe] pelo caderno. (resumir, estudar)

7. Eles levarão **os materiais todos** para a aula de desenho. Eles [escolhe] para a aula de desenho. (levar)

8. Vamos dizer aos primos para virem também? Vamos [escolhe] dizer-lhes. (dizer)

9. Se vieres comigo ao baile, dançarei todas **as valsas** contigo. Se vieres comigo ao baile, [escolhe] todas contigo. (dançar)

10. Na próxima semana não haverá aula. Podes avisar **os colegas**, por favor, João? Sim, eu [escolhe]. (avisar)

Opções para a frase 2:

- [escolhe]
- Dir-lhe-emos
- Fá-lo-ás
- Lê-lo-ei
- avisá-los-ei
- dançá-las-ei
- dizer-lhes
- enviá-la-á
- estudá-la-ei
- levá-los-ão
- resumi-los
- telefonar-lhe-emos

SUBMITER

(a) Multiple-choice example.

Nós fazemos o exercício. —> Nós fazemo-lo.

Escreve, nos espaços em branco, a forma do verbo conjugado com o pronome, conforme o exemplo.

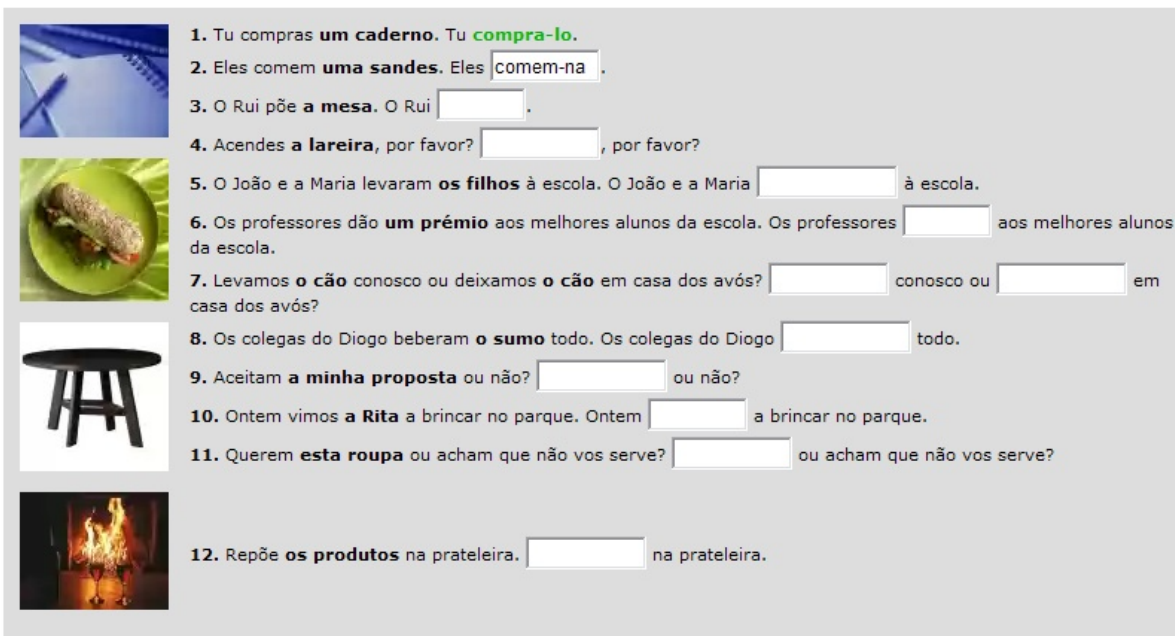
DICA

Nível B1

G

Área: Gramática

Dificuldade: difícil



1. Tu compras **um caderno**. Tu compra-lo.

2. Eles comem **uma sandes**. Eles comem-na.

3. O Rui põe **a mesa**. O Rui [] .

4. Acendes **a lareira**, por favor? [] , por favor?

5. O João e a Maria levaram **os filhos** à escola. O João e a Maria [] à escola.

6. Os professores dão **um prémio** aos melhores alunos da escola. Os professores [] aos melhores alunos da escola.

7. Levamos **o cão** conosco ou deixamos **o cão** em casa dos avós? [] conosco ou [] em casa dos avós?

8. Os colegas do Diogo beberam **o sumo** todo. Os colegas do Diogo [] todo.

9. Aceitam **a minha proposta** ou não? [] ou não?

10. Ontem vimos **a Rita** a brincar no parque. Ontem [] a brincar no parque.

11. Querem **esta roupa** ou acham que não vos serve? [] ou acham que não vos serve?

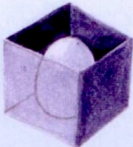
12. Repõe **os produtos** na prateleira. [] na prateleira.

(b) Fill-in-the-blank example.

Figure 2.9: Pronominalization exercises from Ciberescola.

3. Reescreve esta adivinha, substituindo os **pronomes pessoais** destacados pelas palavras que eles representam.

Uma caixa pequenina
mas que pode rebolar,
todos a sabem abrir
ninguém a sabe fechar.



José Viale Moutinho, *Adivinhas Populares Portuguesas*,
6.ª ed., Ed. Notícias, 2000


(Solução na página 83.)

Figure 2.10: Example exercise from *Diálogos 7* (Costa & Mendonça, 2011).



Figure 2.11: Example of incorrect use of pronouns (*achas-te* instead of *achaste*).

Exercise Generation Architecture



The previous exercise generation architecture and its implementation made it difficult to factorize and adapt it to the new exercise that is here proposed. The previous syntactic exercises used cloze questions, in which the distractors are words that fill a blank space replacing the missing word. On the other hand, in the pronominalization exercise, the distractors are sentences built anew by manipulating the syntactic construction of the original stem sentence, namely by deleting and adding lexical material and by changing some of the stem's words (the verb), adjusting it to the pronoun shape (and vice-versa). Instead of just one word to be deleted as in the cloze questions, the constituent to be pronominalized can be made of several words, implying the recursive analysis of syntactic dependencies. The distractor generation has syntactic and positional characteristics that also increase the complexity of the exercise. Finally, the feedback to the students should have complex automatically-generated explanations that use several syntactic features present in sentence.

The following problems needed to be minimized:

Selection rules complexity The sentence selection is complex, involving the analysis of several dependencies, node features, and node order. The constituent selection involves recursive dependency analysis.

Several different sentence types Each sentence type has several complex selection rules, also influencing the distractor generation.

Generation metadata Several syntactic features need to be associated with the questions, in order to be used in the feedback system, and to understand the exercise generation.

The approach presented in the FAST paper (section 2.3.4), to "systematically convert syntactic features into test patterns", can be seen as a useful framework for the automatic generation of syntactic exercises, abstracting the common characteristics of different language topic questions. The ArikIturri (Aldabe, 2011) question model takes this abstraction one step further, and can be a source of good practices.

To tackle these problems, a new architecture was designed. While developed in order to simplify the implementation of the pronominalization exercise, the intention behind this new architecture is that

it be easily applied in the creation of future exercises, so that it may evolve into a framework for exercise generation. The general architecture is presented in Figure 3.1.

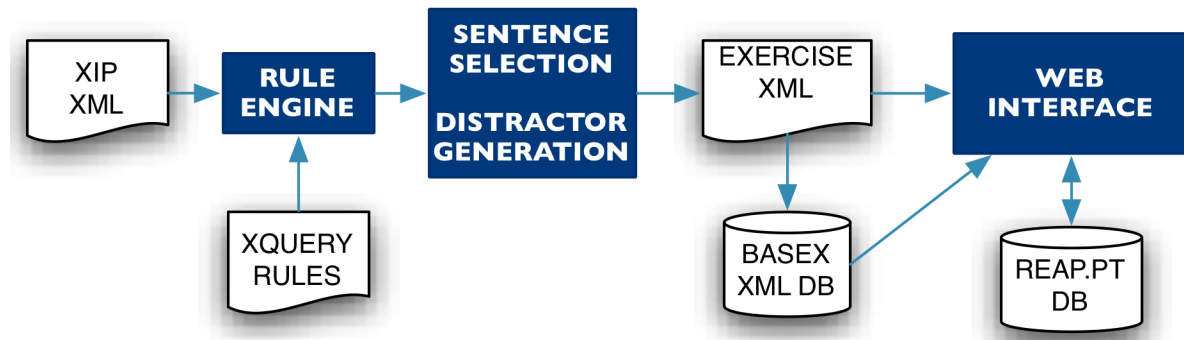


Figure 3.1: REAP.PT new syntactic exercises architecture.

In order to develop the exercises, the STRING (N. J. Mamede et al., 2012) NLP processing chain is used to analyze the corpus sentences, which outputs the syntactic tree and dependencies in XML (N. Mamede et al., 2011). The need for a high-level XML processing language was identified, to replace the existing use of the DOM (see section 2.1.3), one of the leading causes of complexity. In addition, to satisfy the requirement of generation metadata, the exercises themselves are to be generated in XML, making it easier process and add new attributes.

Several alternatives were considered, namely *Scala* (Emir, 2003), *XDuce* (Hosoya & Pierce, 2003), *CDuce* (Benzaken et al., 2003), and *XQuery* (Chamberlin, 2003):

Scala is a general-purpose functional and object-oriented language with native XML support, including pattern matching, literals, and expressions, along with standard XML libraries.

XDuce ("transduce") is a typed programming language (similar to ML), that is specifically designed for processing XML data. It has static typechecking based on regular expression types, and regular expression pattern matching. There is an extension for C#, Xtatic (Gapeyev et al., 2005), providing native XML processing.

CDuce is a general purpose typed functional programming language, whose design is targeted to XML applications. It extends XDuce, introducing less XML specific type constructions (Benzaken et al., 2003).

XQuery is a query and functional programming language that is designed to query collections and do transformations on XML data. XQuery 1.0 was developed by the XML Query working group of the W3C, and became a *W3C Recommendation*. XQuery is a superset of XPath, and uses the path

expression syntax to address specific parts of an XML document. It also has other composable expressions, including an SQL-like "FLWOR expression" for performing joins, literals, comparison (including document order operators), sequence expressions and constructors.

Xquery was ultimately chosen, for several reasons:

W3C Standard Having the *W3C recommendation* makes it a standard that is widely used in many contexts, and the available resources about the language are more widespread than for the other options. This includes learning material, and a higher number of people proficient in the language, possibly reducing learning as well as development effort. The creation of future exercises in the REAP.PT system and its maintenance by new developers was also taken into consideration.

Available implementations Because it is a widely used standard, there are many efficient and free implementations, and integration with popular programming languages, including Java, one of the main languages used in the current REAP.PT system. This also reduces future maintenance effort.

Native databases There are several open-source native XML databases that include XQuery processors. The BaseX¹ XQuery processor was used to process the corpus syntactic analysis and generate the exercises. The Basex database was then used to store the exercises and allows for fast queries and flexibility in the schema, allowing new syntactic information to be easily added as attributes.

Useful and efficient operators While it lacks regular expression pattern matching, XQuery has document order comparison operators and node selection XPath axis, that can be used to analyze some node order features, and that proved to be useful in the sentence selection phase. The union operator, that keeps nodes in document order, was also useful in the generation of answers and distractors. The high-level operators arguably reduced the code complexity and allowed for faster coding (compared to brief experiments with the previous DOM framework).

3.1 Rule Engine

Since the analyzed corpus (with the STRING processing chain) used to generate the exercises is approximately 165GB in size, the Hadoop² Map-Reduce framework for distributed processing was used. It had already been used in the previous syntactic exercises for sentence selection, using the DOM (see section 2.1.3). But this required a new verbose Java program for each exercise, increasing complexity.

A new Java program was created, that uses the Hadoop framework and processes sentences (XML *LUNIT* nodes), using the map function. It searches a *rules* folder for XQuery files, each representing

¹<http://basex.org> (last visited in October 2012)

²<http://hadoop.apache.org> (last visited in July 2012)

a rule that selects and processes a sentence type. Since rules for each sentence type can become quite complex, it is useful to isolate them. Each *LUNIT* node is then processed with each rule, outputting the exercise XML generated from that sentence. Depending on configuration, sentences can be processed by all rules; for example, if the same sentence can produce different exercises using different rules; or if there is rule precedence, they can be ordered alphabetically and only the first successful rule would be used. This is useful when sentences have overlapping features, but one takes precedence over the others, eliminating the need to create rules for all possible combinations of sentence types.

The program can also be used to count sentence-types. In this case, the Hadoop reduce function is used to count equal outputs of the XQuery "rules", that output information about the sentences that should be counted, and text serialization is used instead of XML. Separators between features can be used for later importing and analysis using spreadsheets. This functionality was used in the study presented on section 4.6.2.1.

3.2 *XQuery Rules*

Each XQuery "rule" selects a type of sentence, using several features and dependencies, and generates the exercise according to that sentence type. Some examples are negative sentences, subordinate clauses or the presence of a verbal chain (with auxiliary verb).

Since in the proposed exercise the answer and distractor generation required the analysis of many syntactic features and dependencies, it was done at the same time as the sentence selection (as opposed to the previous exercises, in which all distractors were generated on-the-fly in the interface, since they did not require the sentence analysis). The number of distractors was also limited for each type. When a distractor type does not require the analysis of syntactic information and has many possible variations, it can be generated on-the-fly by the interface (for example, if the variation is in word form).

For the XQuery rules, a module was created factorizing the code common to all sentence-type rules. The rule engine program along with this function module could be used in the development new exercises, and while untested in this regards as only one exercise was developed, could be the beginning of an exercise generation framework. As an example, the functions that output the exercise can receive in their arguments sequences of attributes to be present in the exercise (for example, with features explaining the exercise generation).

One example of an XQuery rule developed for the Pronominalization exercise can be found on Appendix A.

4

Pronominalization Exercise

The goal of this exercise is to substitute a constituent with a pronoun, in a given sentence.

Pronouns can have tonic or atonic forms. Tonic pronouns correspond to nominative forms, appearing as subject independent forms or preceded by a preposition. Atonic forms are prone to cliticization, when they are moved next to a verb and, if after it, connected by a hyphen (-). For this exercise we are interested in the atonic forms, because they are the most problematic to students, since they have more complex restrictions (involving a high number of features and dependencies).

The list of atonic pronouns is: *me, te, se, nos, vos* / *o, a, os, as* / *lhe, lhes*. Only the 3rd person pronouns will be considered, because those are the ones that can substitute a complement in the accusative or dative cases.

There are three grammatical aspects present in pronominalization exercises that are interconnected:

Form The form of the pronoun, according to the verb termination, and the spelling rules of the verb. Contractions of two pronouns also have to be considered.

For example, if the verb terminates with *-r, -s* or *-z*, the accusative, 3rd person pronouns *o, a, os, as* assume the form *lo, la, los, las*. In that case, the verb loses its last letter and it is accentuated according to general spelling rules. If the verb terminates with nasal sounds *-m, -õe* or *-ão*, the same pronouns assume the form *no, na, nos, nas*, but the verb remains unaltered.

Case The case of the pronoun, according to its syntactic function. The complement function is determined by the verb it depends of and the pronouns that replaces it takes the correspondent case, as presented on Table 4.1.

Position The position of the pronoun in the sentence. It can appear at the left or right of the verb. In the future or conditional tenses, it appears between the verbal root form and the tense ending

Table 4.1: Pronominal case in Portuguese

Case	Syntactic Function	Constituent Form	Atonic Pronouns
dative	indirect complement	prepositional phrase	me, te, lhe, nos, vos, lhes
accusative	direct complement	noun phrase	o, a, os, as
accusative	direct complement	substantive subordinate clause	o (invariable)
oblique	-	prepositional phrase	mim, ti, si (tonic)

morphemes (*lavá-lo-ei* “I will wash it”; *lavá-lo-ia* “I would wash it”). If the pronoun happens to be found after the verb, an hyphen should be used between the verb and the pronoun. The rules governing the placement of clitics are very complex and even native speakers have a problem to do it correctly, this being a major fracturing phenomenon of the language (Móia & Peres, 2003). The main factors involved are: the verb is in a subclause, or under a negation; the presence of auxiliaries and the nominal form of the main verb they are construed with (infinitive, gerund or past participle); the indefinite or negative type of subject; the presence of adverbs before or after the verb, etc. These factors are detailed in section 4.6.

These three aspects are interconnected, but can, in some cases, be presented to the student independently by the design of the exercise; for example, presenting distractors that vary only in one of these aspects, so as to teach these aspects to the student gradually and according to his/her proficiency level.

4.1 Examples

4.1.1 Accusative case

Choose the right pronominalization of the constituent signaled in bold:

- Stem from the corpus:
- *O Pedro deu o livro à Ana.* (Pedro gave **the book** to Ana.)

Correct answer:

- *O Pedro deu-o à Ana.* (Pedro gave her **the book**.)

[The pronoun is in the accusative case because the constituent **o livro** (the book) is the direct complement of the verb **deu** (gave). The correct position for the clitic is after the verb, so a hyphen should be used.]

Distractors:

- *O Pedro deu-lhe à Ana.* (Pedro gave **to_him** to Ana.)

[Dative case instead of accusative.]

- *O Pedro deu-lo à Ana.* (Pedro gave **it** to Ana.)

[Wrong choice of pronoun form, considering the verb termination.]

- *O Pedro o deu à Ana.* (Pedro **it** gave to Ana.)

[Wrong clitic position.]

4.1.2 Dative case

Choose the right pronominalization of the constituent signaled in bold:

- Stem from the corpus:
- *O Pedro deu um livro à Ana.* (Pedro gave a book **to Ana**.)

Correct answer:

- *O Pedro deu-lhe um livro.* (Pedro gave **her** a book.)

[The pronoun is in the dative case because the constituent **à Ana** (to Ana) (prepositional phrase) is the indirect complement of the verb **deu** (gave). The correct position for the clitic is after the verb, so a hyphen should be used.]

Distractors:

- *O Pedro deu-a a um livro.* (Pedro gave **her** to a book.)

[Accusative case instead of dative. The valence or syntactic construction of the verb is also important to generate a convincing distractor; in this case, *the book* was rephrased as a prepositional complement, so that the verb could keep both direct and indirect complement, instead of just producing the wrong form, as in *O Pedro deu-a um livro.*, which is less convincing.]

- *O Pedro deu um livro a ela.* (Pedro gave a book to her.)

[Oblique case; “wrong”, or less canonical, use of a tonic form, instead of the dative (atonic).]

- *O Pedro deu-la um livro.*

[Wrong choice of pronoun form, considering the verb termination.]

Other difficulty levels:

In further versions of the exercise, tonic pronouns could also be considered, in particular the most difficult forms, like possessive and oblique pronouns.

- *O Pedro leu o livro da Ana.* (Pedro read **Ana’s** book.)
= *O Pedro leu o seu livro.* (Pedro read **her** book.)

= O Pedro leu o livro *dela*. (Pedro read **of her** book.)

Considering auxiliary verbs, future and conditional tenses can also make more advanced exercises for the student, but this would also make them more complex to generate because of the additional rules and complex inter-relations.

4.2 Specific Exercise Architecture

For this exercise, the rule engine program was used to process the sentences with several XQuery “rules”. One rule was used for each set of sentence features that affect the complement to be pronominalized. These rules are associated with the pronoun positioning rules (loosely referred to as *sentence types* in this document). This allows to better isolate the sentence type selection that affects clitic positioning, since it is a major linguist problem and the most complex for this exercise, involving the higher number of features and dependencies (refer to section 4.6).

One example of the exercise output for one sentence is presented in Listing 4.1.

Listing 4.1: Pronominalization exercise output example.

```
1 <LUNIT start="43927" end="44020">
2   <original auxcase="1" clause="POS" comp="os investidores de a Bolsa de Zurique"
3     file="/corpora/publico/20121004/Parte15/Parte15adq.out"
4     prep="a" rule="2"iaux="está" verb="animar">
5     A recuperação do dólar face ao franco suíço está a animar [[ os investidores da Bolsa de Zurique ]].
6   </original>
7   <answer>
8     <response accusative="true" aux="false" position_after="true">
9       A recuperação do dólar face ao franco suíço está a animá-{{los}}.
10    </response>
11    <response accusative="true" aux="true" position_after="true">
12      A recuperação do dólar face ao franco suíço está-{{os}} a animar.
13    </response>
14  </answer>
15  <distractors>
16    <response accusative="false" aux="false" position_after="true">
17      A recuperação do dólar face ao franco suíço está a animar-{{lhes}}.
18    </response>
19    <response accusative="false" aux="true" position_after="true">
20      A recuperação do dólar face ao franco suíço está-{{lhes}} a animar.
21    </response>
22    <response accusative="true" aux="false" position_after="false">
23      A recuperação do dólar face ao franco suíço está a {{os}} animar.
24    </response>
25    <response accusative="false" aux="false" position_after="false">
26      A recuperação do dólar face ao franco suíço está a {{lhes}} animar.
27    </response>
28  </distractors>
29 </LUNIT>
```

In the *<original>* element of this listing there are attributes of the target sentence, such as clause type (POS=affirmative), the complement, main and auxiliary verbs, linking preposition, rule number, and corpus file path. On the sentence, the complement is enclosed in brackets. The *<answer>* element contains the correct answers (*<response>* elements), and the *<distractors>* element has the wrong answers. Each *<response>* element also includes metadata such as pronoun case, position, and verb (main or auxiliary).

In this example, the second answer has a clitic positioning error related to infinitive direct complements (refer to section 6.2). However, only the first answer is presented to the students, being the most canonical, thus this error does not affect the exercise in practice. Nevertheless, this error is caused by the incorrect encoding of clitic positioning with auxiliary verbs, and can easily be corrected by changing the corresponding tables.

Each sentence could in principle be selected by more than one rule, for two reasons:

- Each sentence can have several complements that can be pronominalized, thus generating more than one exercise. The complements can be in different clauses, and so can be affected by sets of features belonging to different rules / sentence types. In this case, each complement is processed by the corresponding rule and ignored by the others.
- It is possible that more than one rule applies to a single complement, because the feature sets can overlap. For example, a negative clause that attracts the clitic to the pre-verbal position, and a clitic-attracting adverb after the verb. This combinations complicate the exercise both in terms of coding and to the student, so they were not explored in the present work. Since the rules are complex, it is arguably better to teach them to the students separately and not in combination. The rules are therefore coded as mutually exclusive, eliminating sentences with complements in clauses that are affected by multiple rules. However, solutions to this problem were considered. In this case, most of the combinations can be solved by setting rule precedence, which can be done in the rule engine program, by ordering the rules names alphabetically. The rules would cease to be mutually exclusive, and when a rule were matched, the others would be discarded. This feature can be used in future exercises that may require it, or to teach the precedence of the clitic positioning rules.

4.3 Sentence Selection

There are two approaches when it comes to choosing a stem from the corpus. It can be a sentence that already has the pronouns replacing an antecedent, or a sentence that does not have pronouns. In the first approach, the sentence represents the correct pronominalization answer of the exercise, so we have

to generate the question sentence and the distractors. To generate the question sentence, the antecedents need to be found, and the right position on the sentence to put them. On the other hand, if we start with a sentence that does not have pronouns, a constituent has to be selected to generate the target sentence (correct answer). In both options, distractors must also be generated. The second approach was chosen, because on the first there is the problem of anaphora resolution to find the antecedent, which is a difficult problem, as noted section 2.4. In this approach, the generation process starts with a sentence from the corpus, from where target patterns (constituents) are extracted.

Several filters were added to eliminate unsuitable exercises. The first are sentence filters that apply to all sentences and rules:

Word number Sentences can be filtered when they exceed a maximum number of words, to make them simpler for the students and less prone to NLP errors.

Pronoun case For accusative case exercises, sentences which already have accusative clitics in the third person are discarded, because the student could deduct the correct answer from examples on the sentence (e.g. *O Pedro encontrou a Ana e cumprimentou-a alegremente.*). The same is done for dative case exercises.

There are also filters to prevent sentences with NLP analysis errors to be proposed for generation. One such filter is done on affirmative main clauses, to eliminate sentences that are in reality subordinate clauses that were incorrectly analyzed. It was noted that many of those sentences had the ambiguous word *que*¹ before the complement, which many times introduces a subordinate clause, but was analyzed as a main clause. If all sentences with *que* before the complement were filtered, the relative may not have been affecting the complement, and too many correctly analyzed sentences would be discarded. To solve this problem, the sentences were only filtered when the relative was not separated from the complement by a punctuation mark. While this solution is not linguistically perfect, it proved to work well and filter many incorrectly analyzed sentences. One example taken from the corpus is the sentence “*Afinal, parece que consumir fura os tímpanos.*” (After all, it seems that consuming pierces the eardrums). In this sentence, the *que* introduces a subclause. However, in the incorrect analysis, the verb *fura* that affects the complement is not in the subordinate clause, because there is no dependency connecting it to the previous verb *consumir*. If the main clause rule were used, the clitic would be in the post-verbal position, yielding a wrong answer *Afinal , parece que consumir fura-os.*, instead of the correct pre-verbal position *Afinal , parece que consumir os fura.*, using the subordinate rule (refer to section 4.6).

Other filters apply to each phase of the generation, and are described in the following sections.

¹In European Portuguese, *que* can be the subordinative conjunction (that), the relative pronoun (that, which), the interrogative pronoun (what, which), and even a linking word in the auxiliary verbal chain *ter que* +infinitive (have to).

4.4 Complement Selection and Analysis

The pronoun case is an argument of the rules, and it is used to get the complement dependencies corresponding to the accusative (“*CDIR*” dependency) or dative (“*CINDIR*”) cases.

In the evaluation, only the accusative case was tested, using the direct complement dependency, because the indirect complement dependency was not present in enough sentences in testing, and because it is not fully implemented in the STRING processing chain yet. In the first 2000 sentences of the CETEMPúblico corpus used for development testing, only 15 had the *CINDIR* dependency, and of those, none passed the pronominalization filters. However, the dative case pronominalization was implemented for the most part, since most of the code is generic, and the positioning rules are almost the same (cf. section 4.6).

The following filters were applied to the complement selection²:

Noun phrases Complements have to be noun phrases in which the head is a noun. An exception occurs when the complement dependency is in a prepositional phrase and has a determiner quantifier (*QUANTD* dependency), as seen in the example *Recusou, contudo, as propostas dos governos estrangeiros de enviar para o país **equipas de especialistas***, in which *especialistas* is marked as the direct complement, and *equipas* its quantifier determiner (see section 2.1.2 for a definition).

Subclauses When the direct complement is a subclause (subordinate completive clause), it has the *SENTENTIAL* dependency, and should not be pronominalized in this exercises.

Indefinite complements Indefinite complements cannot be pronominalized. The complements must have a determiner (*DETD* or *QUANTD* dependencies), and the determiner cannot have the *INDEF* feature.

Appositions The complements cannot have appositions (ex: *A UNITA declarou ontem que o seu líder, Jonas Savimbi, não aceitará o cargo de vice-presidente que lhe é proposto nos acordos de paz de Angola*), nor can they be appositions of another noun phrase (ex: *Raisa Gorbatchov ganha Donna– O prémio literário Donna- Cidade de Roma, 1992, foi atribuído a Raisa Gorbatchov, pela sua biografia-testemunho, « Io Spero », editada no Verão passado, anunciou ontem, na capital italiana, **a presidente do júri**, a escritora Gabriella Sobrino.*); so they cannot be in any argument of the *APPOSIT* dependency.

Relatives If there is a relative clause after the complement, introduced by a prep *que/o qual/cujo*, the complement cannot be pronominalized (ex: *Os assaltantes atraíram **a atenção de uma de as funcionárias**, que deu de imediato o alarme.*).

²The examples were taken from the corpus

The complement dependencies in STRING only detect the head of the constituent. To recover the entire constituent, several steps were taken, some of which may not be linguistically correct in every sentence, for lack of linguistic information in the analysis. Some decisions are taken on a best effort basis, where the number of correct pronominalizations is believed, in an educated guess, to outnumber the number of errors that are introduced.

For example, when there is a conjunction of several complement dependencies on the same verb, they are joined, and the constituent is considered to span every word from the first to the last complement in document order. This happens because the XIP-PT dependencies are binary by design. In the sentence *A Ana comeu a banana deliciosa e a suculenta maçã.* (Ana has eaten the banana and the apple), the verb *comeu* (has eaten) has two CDIR dependencies, *banana* and *maçã*.

The basic selection consists of including the whole node in which the complement head appears (usually a noun phrase, but can be a prepositional phrase in case it is preceded by a determiner quantifier). Then, for each complement head, modifiers are added in a recursive fashion. The modifiers can be adjectives or prepositional phrases which start with *de* (or). If a proper noun immediately follows (without punctuation) the whole complement, it is also added, since there is a very high probability of belonging to it. The modifiers can only be included in the complement if they immediately follow it (ignoring punctuation and conjunctions, as in *os próximos ministros de a Defesa e de as Relações Exteriores*), since there can be adjective modifier dependencies that apply to the complement head that are separated from it and do not belong to the constituent. And there can be recursive modifiers to the modifiers, which must also be included. This is why the attachment must be done in a recursive and incremental method.

In the sentence *A GF confiscou ainda a viatura ligeira de marca Bedford.*, the PP *de marca* was added because it starts with *de*, and *Bedford* was added for being a proper noun that follows the complement.

When a PP is attached to the complement incorrectly, or when a PP should be part of the complement but is not for lack of linguistic information, the well-known *PP-attachment* problem occurs. This problem cannot currently be solved using the information provided by the STRING processing chain. The first case can be exemplified in the sentence *Importante é acima de tudo a noção de servir [o utente de forma] eficaz.*, in which the PP should not have been included in the complement. The second case can be seen in the sentence *As exportações serviriam para justificar [a saída dos materiais] comprados por Joaquim Oliveira.*, in which the last PP was not attached to the complement as it should.

4.4.1 Gender and Number Selection

In order to be pronominalized with correct agreement, the gender and number of the complement need to be calculated. In principle, the gender and number of the head of the complement are used for this

calculation. If the determiner is an article, its gender/number are used. And if there is a determiner quantifier, the decision depends on its partitive nature. If the quantifier is partitive (*SEM-MEASOTHER* feature), the gender/number are that of the complement head (ex: *metade do investimento total*, pronoun: *o*). Otherwise, the gender/number comes from the quantifier (ex: *fardos de palha*, pronoun: *os*).

If there is more than one complement head, the number is plural, and the masculine gender takes precedence over the feminine (e.g. *O João levou a Teresa e o Carlos ao cinema.*, becomes *O João levou-os ao cinema.*).

4.5 Pronoun Case and Form Generation

As mentioned above, the case is an argument of the generation and depends on the complement dependency. In the dative case, since only 3rd person pronouns were considered for this exercise, only two are used which differ in number. In the accusative case, the pronouns are selected in agreement with gender and number, using a map. However, when they occur connected to the verb by an hyphen, they assume different forms. A function calculated the right form according to the basic accusative pronoun and the verb termination, additionally changing the verb termination according to spelling rules.

4.6 Pronoun Positioning Rules

The clitic positioning rules are presented in detail in a working paper by Baptista (Baptista, 2012). A summary is provided here for reference.

There are 6 rules for complement pronouncing, common to both accusative and datives pronouns. As explained in section 4.2, they correspond to the 6 XQuery rules that isolate their complexity, and are an important distinguishing factor between the generated exercises.

All rules record common generation information (eg. for feedback purposes), namely the verb, it's complement, the pronominalized pronoun along with case and position, rule number, original file and offset in the corpus. Specific additional information is mentioned bellow in each rule.

When the verb is in the future-indicative tense or in the conditional, and the clitic position is after the verb, the clitic is placed between the thematic vowel and the verb tense endings; this phenomenon is called mesoclis, and was not implemented in the present work, although it is trivial to implement given the modular nature of the code. Complements with verbs in such tenses are thus not used for generation.

The dative clitic positioning presents the same general constraints as the accusative (Baptista, 2012), except for verbal chains, in which unlike the accusative, the dative pronoun alone can be fronted and

attached to the auxiliary verb. There are also some few differences in the presence of some indefinite subjects.

4.6.1 Rule 1: Simplest case of affirmative main clauses without verbal chains

The clitic is placed after the verb and linked by an hyphen, if the verb is the main verb in an affirmative clause; this phenomenon is called enclisis.

This case can be seen in the following example taken from the corpus:

Mário Soares, por seu lado, elogiou a personalidade do visitante..

Mário Soares, por seu lado, elogiou-a.

4.6.2 Rule 2: Verbal chains

This is the most complex rule, since the constraints are different for each auxiliary verb, and there are many possible variations, depending on the presence of negation, insertion in a subclause and linking preposition.

In this exercise only verbal chains with one auxiliary verb are considered, in order to simplify the students learning and the exercise generation rules.

There can be four possible positions:

- The clitic is attached to the main verb (enclisis);
- the clitic is moved to the front of the main verb (proclisis);
- the clitic is attached to the auxiliary verb (enclisis);
- the clitic is moved to the front of the auxiliary verb (proclisis).

Only the first tree apply to main clauses, while all four can apply to subclauses and in negative sentences, giving a total of 12 combinations of sentence types and positions.

In his paper, (Baptista, 2012) provides an appendix with two tables detailing the clitic positioning within auxiliary verb chains with and without linking preposition. For each auxiliary verb, preposition, verb-form, gram-value, and type, it is shown for which of the 12 possible combinations of position/sentence type the pronominalization is correct or incorrect. There can be more than one correct position for each verb and feature set.

The features are the following:

preposition designates the preposition linking the auxiliary to the main verb;

verb-form indicates the non-inflected form of the main verb;

gram-value temporal, aspectual and modal grammatical values conveyed by the auxiliary, which were later found not to be needed to distinguish the auxiliary verbs for clitic positioning purposes.

type can be: VASP: aspectual auxiliary verb; VMOD: modal auxiliary verb; VTEMP: temporal auxiliary verb.

The tables were directly implemented using a map, using as key the concatenation of auxiliary verb, their distinguishing features, and sentence features (main affirmative, main negative or subclause), mapped to the four possible positions (further divided in correct and incorrect positions sets, respectively for the answers and distractors).

All the features were recorded as attributes in the exercise output, for generation information used in the feedback interface.

4.6.2.1 Clitic Positioning within verbal chains: Empirical Study

The data used in the clitic positioning paper (Baptista, 2012) “was obtained by introspection alone”, using example sentences to derive the correct positioning for each feature set. It has an appendix which tries to systematically present the constraints on accusative clitic positioning within verbal chains.

However, given the complexity of the positional constraints, an introspective experimental protocol alone may not be enough to guarantee a high level of confidence in agreement with real language use.

As such, a study using the corpus and the STRING NLP processing chain was performed in this work, counting the number of occurrences of clitic positions in each of the auxiliary verbs and recording the presence of the same features used in the introspective study. For practical reasons, for the counting of the clitic occurrences, only the last auxiliary (i.e. the one before the main verb) was considered in longer verbal chains (two or more auxiliaries). Though pending a detailed study it is noteworthy that, for the most part, results seem to confirm the introspective definition of positioning rules.

The study results can be found on Appendix B.

4.6.3 Rule 3: Clitic attraction by negation

In negative sentences with negation adverbs *não* ‘no/not’, *nunca/jamais* ‘never’, *nem* ‘not even/nor’, and the like, the clitic is attracted to the pre-verb position.

The negation is checked by looking at the *NEG* feature in the verb modifier dependencies **MOD**.

This case can be seen in the following example taken from the corpus:

Não copiamos os nossos vizinhos, mas tentamos ser um exemplo.

Não os copiamos, mas tentamos ser um exemplo.

4.6.4 Rule 4: Indefinite and negative subjects

This rule deals with pronouns and determiners that modify the subject.

Indefinite pronouns, e.g. *alguém* ‘somebody’, *algo* ‘something’ *tudo* ‘everything’ and negative indefinite pronouns e.g. *ninguém* ‘nobody’, *nada* ‘nothing’, attract the clitic pronoun to the pre-verb position.

This also happens when the subject is a common noun with some quantifier determiners, e.g. *todos* ‘all’ or *ambos* ‘both’, and some indefinite determiners, e.g. *algum* ‘some’ or *qualquer* ‘any’.

However, some of these pronouns and determiners allow both clitic positions, and so don’t generate position distractors.

The subject itself can also be one of these pronouns, instead of being modified by one, as seen in the following examples:

- *Todos os rapazes jogam à bola.* (quantifier determiner *todos* modifying the subject *rapazes*).
- *Todos jogam à bola.* (the subject is the quantifier determiner).

The *DETD* and *QUANTD* dependencies on the subject head were used to get these pronouns. In order to differentiate between them, both for positional and feedback purposes, specific lists were used, since the features from the analysis were not conclusive to determine the type: indefinite pronouns, indefinite determiners, and quantifier determiners.

The pronoun and its type were recorded as attributes in the exercise output, for generation information used in the feedback interface.

4.6.5 Rule 5: Clitic-attracting adverbs

Adverbs allowing both pre- and post-verbal position, attract or leave clitic in its basic position, respectively, depending on the position they occupy in the sentence in relation to the verb they modify.

When there are both pre- and post-verbal clitic-attracting adverbs, the clitic position in the right answer defaulted to the post-verbal position (enclisis), since it is the general position in affirmative main clauses. When this default happens, the position distractor is not presented. As mentioned before, rule combinations are not currently generated. If combinations were used, negation would take precedence over clitic-attracting adverbs (in a negative sentence with an adverb in the post-verbal position).

The clitic-attracting adverb was recorded as an attribute in the exercise output, for generation information used in the feedback interface.

4.6.6 Rule 6: Subordinate clauses

In subordinate clauses, clitics are attracted to pre-verbal position. This takes place in completives, relatives and adverbial subordinate clauses.

In subordinate adverbial infinitive clauses introduced by the subordinative conjunction *ao* ‘to-the’ (e.g. “*A Ana descobriu isso ao ler o jornal.*”), the subordinate status is ignored for clitic positioning purposes. This is detected by looking for an *INTROD* dependency, which links the first element of the SC chunk to the head of the main verb of the subclause (N. Mamede et al., 2011), and checking if it has the *TEMPORAL* feature and the *ao* lemma.

4.7 Response Generation

When generating the responses, both the correct sentences and the distractors, if a verb is capitalized in the original sentence and the pronoun is inserted in the pre-verbal position, the verb needs to be uncapitalized and the pronoun capitalized. This phenomenon can be seen in the distractor of the following example from the corpus: “*Ganhei a medalha de prata*”, diz-lhe ofegante.: “*A ganhei*”, diz-lhe ofegante.. The correct answer for this example is “*Ganhei-a*”, diz-lhe ofegante..

Contractions between pronouns are also dealt with using substitutions. In the pronoun generation, if there was already a pronoun, both would be left in their correct positions. The pair is thus substituted by the correct contraction (27 different contractions were considered). This can be seen in the corpus sentence *Pouco depois, o príncipe aponta-lhe a arma ao ventre.*, where the correct pronominalization is *Pouco depois, o príncipe aponta-lha ao ventre.* (contraction of the dative pronoun *lhe* with the accusative *a*).

The XQuery union and node order operators were useful to remove the previously calculated constituent nodes from the sentence, and insert the pronoun in the correct position. The high level and versatility of the XQuery language proved to be a good choice to simplify the development and minimize errors in the generation.

4.7.1 Distractor Generation

There are four types of distractors:

- Wrong case distractors

- Wrong position distractors
- Combinations of wrong case and position
- Wrong accusative form distractors

The gender and number are always kept in agreement with the right answer, because making them vary in a distractor would result in a too obvious exercise.

The case and position distractors are generated by the same function that generates the correct answer, by changing the arguments of the case and position. This is done during the generation phase, since their number is low enough, the processing is complex and uses the analysis information that is already available in memory from the answer generation.

However, the accusative case form distractors, in which the form is incorrect, can be easily generated from the correct answer by the removal or addition of one character in the clitic. The number of possible combinations makes it easy to randomly generate those distractors during the exercise presentation, saving space in the exercise storage.

4.8 *Exercise Interface*

4.8.1 **Question Interface**

In the question interface, the original sentence, correct answer and distractors are presented to the student as a multiple-choice selection. Four options are always presented, the correct answer and three types of distractor (configurable number), randomly chosen and shuffled.

A button is present for the student to indicate he/she thinks the exercise has errors, in order for the flagged exercises to be examined by the teacher later.

An example of the question interface can be seen on Figure 4.1.

4.8.2 **Feedback Interface**

In the generation, syntactic information about the exercise generation is stored. It is then used in the interface to present feedback about the correct answer to the student, so he/she can understand and learn all the aspects pertaining to the pronominalization (case, position and form), even if the provided answer was correct.

The sentences corresponding to each feedback section and variations are stored, with name suffixes that correspond to exercise attributes (some are booleans, for example `position_after` and `aux`, that tell

Pronominalização

Se substituir a expressão assinalada a negrito por um pronome, qual das frases seguintes é a correta?

*A recuperação do dólar face ao franco suíço está a animar **os investidores da Bolsa de Zurique**.*

- ☐ *A recuperação do dólar face ao franco suíço está-**lhes** a animar.*
- ☒ *A recuperação do dólar face ao franco suíço está a animá-**los**.*
- ☐ *A recuperação do dólar face ao franco suíço está a **os** animar.*
- ☐ *A recuperação do dólar face ao franco suíço está a animá-**os**.*

Responder

Exercício tem erros

Figure 4.1: Exercise question interface.

the pronoun position relative to the verb, and if the verb is main or auxiliary). Some sentences are fixed for every feedback page, while others are retrieved by their names, such as position rule or the attribute suffixes.

Several grammatical explanations are also included in tool-tips that appear when the user hovers the mouse cursor over the underlined words. They allow the students to understand the grammatical concepts in the answer feedback.

In the explanations and inside parenthesis, words taken from the sentence can be seen alongside each mentioned category. This is achieved using a basic template system. Sentences corresponding to each feedback section are stored with placeholders for each category to be replaced by words from the exercise. They are replaced by the correct words taken from exercise attributes during the feedback presentation.

Twenty-two different sentences can be combined to generate the feedback, and each has several placeholders for examples taken from the exercise sentence, and 7 possible tool-tips with grammatical explanations.

An example of the feedback interface can be seen on Figure 4.2, and an additional example including a tool-tip with grammatical explanations on Figure 4.3.

Pronominalização

A sua resposta anterior estava **correta** .

A pergunta anterior era:

Se substituir a expressão assinalada a negrito por um pronome, qual das frases seguintes é a correta?

A recuperação do dólar face ao franco suíço está a animar **os investidores da Bolsa de Zurique**.

A sua resposta foi:

A recuperação do dólar face ao franco suíço está a animá-**los**.

A expressão assinalada a negrito (os investidores de a Bolsa de Zurique) é o complemento direto do verbo (animar).

O pronome que substitui o complemento direto deve estar no caso acusativo.

A posição correta do pronome é ligado por hífen ao verbo principal (animar), já que se trata de uma frase afirmativa e o verbo (animar) apresenta-se construído com verbos auxiliares.

A forma correta do pronome é "**los**", pois o verbo ao qual se liga (animar) termina em -r, -s ou -z.

Próxima Pergunta

Exercício tem erros

Figure 4.2: Exercise feedback interface.

Pronominalização

A sua resposta anterior estava **correta** .

A pergunta anterior era:

Se substituir a expressão assinalada a negrito por um pronome, qual das frases seguintes é a correta?

A recuperação do dólar face ao franco suíço está a animar **os investidores da Bolsa de Zurique**.

A sua resposta foi:

A recuperação do dólar

A expressão assinalada a negrito (os investidores

O pronome que substitui o com

O **complemento direto** é um constituinte essencial de uma construção verbal e que se liga diretamente, isto é, sem preposição, ao verbo de que depende (por exemplo, na frase: *O Pedro leu um livro*, o grupo nominal *um livro* desempenha a função de complemento direto do verbo *leu*).

A posição correta do pronome é ligado por hífen ao verbo principal (animar), já que se trata de uma frase afirmativa e o verbo (animar) apresenta-se construído com verbos auxiliares.

A forma correta do pronome é "**los**", pois o verbo ao qual se liga (animar) termina em -r, -s ou -z.

Próxima Pergunta

Exercício tem erros

Figure 4.3: Exercise feedback interface with tool-tip on mouse-hover.

5 Evaluation

5.1 *Evaluation Setup*

The exercises were generated from the CETEMPUBLICO corpus, “a 180-million word newspaper corpus free for R&D in Portuguese processing.” (Santos & Rocha, 2001), that includes approximately 8 million sentences, according to its official website ¹.

Only sentences with less than 20 words were used for this evaluation, because longer sentences would be more difficult for the students to read, and increased the probability of NLP analysis errors in the STRING processing chain.

Table 5.1 shows the number of exercises that were generated from the corpus, for each sentence type rule (refer to section 4.6). Table 5.2 shows the counts for sentences with less than 20 words.

Table 5.1: Total number of generated exercises

<i>Rule</i>	<i>Exercises #</i>
1 (main clauses)	580,546
2 (verbal chains)	158,939
3 (negation)	28,653
4 (indefinite subjects)	11,533
5 (adverbs)	107,779
6 (subclauses)	405,438
Total	1,292,888

5.1.1 Expert Analysis

The evaluation of exercises generated from the corpus cannot encompass all generated exercises. On one hand, as the number of generated exercises is too large for manual inspection, even considering only the number of sentences with less than 20 words (almost 207k), determining the total number of correct exercises is not trivial. On the other hand, other factors may complicate the matter further, as the number of possible solutions and the number of distractors for a single stem may vary. Besides

¹<http://www.linguateca.pt/CETEMPUBLICO> (last visited in October 2012).

Table 5.2: Number of generated exercises for sentences with less than 20 words

<i>Rule</i>	<i>Exercises #</i>
1 (main clauses)	100,706
2 (verbal chains)	29,926
3 (negation)	6,390
4 (indefinite subjects)	2,590
5 (adverbs)	17,878
6 (subclauses)	49,486
total	206,976

these aspects, in some sentences there are more than one possible target complements that can undergo pronominalization, hence the number of possible exercises for a single source sentence can become relatively large. Furthermore, it is necessary, in exercises error analysis, to distinguish between errors due to the generation process and those errors due to the previous NLP steps.

For the above reasons, an expert linguist analyzed a random sample of exercises generated from the whole corpus. The exercises were classified by grammatical correction, and annotated with error cause classes. Each exercise can be annotated with more than one error.

Two random samples of 120 exercises were retrieved (20 for each of the 6 rules), giving a total of 240 exercises. The samples were shuffled, and all attributes with information from the generation were stripped (including the rule), in order to remove the bias that the NLP analysis could introduce. The whole information was then used to determine the cause of the eventually incorrect exercises.

5.1.2 Expert Evaluation Measures

Precision was defined as the number of correct exercises by the total number of evaluated exercises.

Recall could be defined for this problem as the number of correct exercises by the total number of possible exercises (correct + missed). However, this calculation cannot be performed with the random set of exercises generated for this evaluation. It would be more adequate to start from a set of sentences, and then proceed with the exercises definition, against which the exercises generated by the system would finally be compared. The problem that arises is that it is not trivial to determine a priori the number of possible exercises (possible pronominalizations) that can be generated. Besides, as noted in the previous section, there are more than one possible target complements that can undergo pronominalization in a sentence. This work attempts to generate exercises from most sentences that have the complement dependencies, because the reasons for the possibility that a complement can undergo pronominalization are not yet completely defined. A possible approach would be to create a set of sentences, and have an expert analyze the number of possible pronominalizations. However, attending

to the pedagogic purposes of this work and the fact that enough exercises are generated for those purposes, the effort associated with that analysis would not be worthwhile. It has been decided, then, that only *Precision* should be accessed at this time, in view of the effective use of the exercises in a real-life ICALL context. In future work, these considerations on Recall calculation may be undertaken.

5.1.3 Crowd-sourced Testing

A website was made available for testing by both native speakers and non-native Portuguese students. Native speakers were used because the exercise difficulty is high enough to be a challenge even for natives, and to analyze agreement with the expert analysis in error detection, since the users were given the option to signal that the presented exercises had errors.

Six randomly chosen exercises were presented to each user, one for each rule that governs clitic choice and positioning (refer to section 4.6). For non-native students, only the exercises deemed correct in the expert analysis were shown, in order not to confuse them or diminish the learning potential. If one exercise was deemed incorrect by the user, a new one of the same rule was presented.

One of the factors to be analyzed was the nature of the errors that are committed by speakers of different levels, namely the distractor type in the wrong answers.

The evaluation website introduction can be seen in Figure 5.1, and the user form in Figure 5.2.

5.1.4 Questionnaire

In the end of the crowd-sourced testing website, a usability and user satisfaction questionnaire was done, in order to identify aspects that could be improved. Some of the statements were based on the standard *USE Questionnaire*². USE stands for Usefulness, Satisfaction, and Ease of use. The questionnaire was constructed as five-point Likert rating scales (a psychometric scale). Users were asked to rate agreement with the statements, ranging from strongly disagree to strongly agree.


The questionnaire was composed of the following statements, plus a free-form commentary text box:

- *O sistema é fácil de utilizar.* (The system is easy to use. - *Ease of Use* USE Factor)
- *Percebi rapidamente o objetivo.* (I understood the objective quickly. - based on *Ease of Learning* USE Factor)

²http://www.stcsig.org/usability/newsletter/0110_measuring_with_use.html (last visited in October 2012)

Pronominalização

Bem vindo!



Obrigado por testar o exercício de **Pronominalização** do **REAP.PT**!

A sua colaboração é muito importante para nós.

Este conjunto de 6 exercícios deverá demorar cerca de **10 minutos**.

O exercício consiste em substituir o complemento directo de um verbo pelo pronome adequado, como no exemplo seguinte:

O Pedro leu **o livro**.
Resposta:
O Pedro leu-**o**.

Por favor complete todos os exercícios, sem recarregar a página nem retroceder.
No final, preencha o questionário de satisfação.
Obrigado.

INICIAR

Figure 5.1: Exercise evaluation website introduction.

- *Os exercícios são demasiado fáceis.* (The exercises are too easy.)
- *O feedback apresentado é suficiente.* (The presented feedback is sufficient.)
- *O sistema é útil: aprendi alguma coisa ao usá-lo.* (The system is useful: I learned something while using it. - *Usefulness* USE Factor)
- *Apreciação global do sistema.* (Global system appreciation. - based on *Satisfaction* USE Factor, uses satisfaction scale instead of Lickert agreement scale)

The questionnaire web-page can be seen in appendix C.

Pronominalização



Antes de iniciar o exercício, por favor preencha o seguinte formulário.

Nome:

E-mail:

Idade: * anos

Língua Materna: *

(se outra, qual?) , com anos de contato com o português.

* - campo de preenchimento obrigatório

Figure 5.2: Exercise evaluation website user form.

5.2 Expert Analysis Results

From the 240 manually analyzed exercises, 75 were found to have errors, and 165 were considered correct. Therefore, the system precision in this evaluation was 68.8%.

As it will be seen bellow, significant percentage of the errors are related to shortcomings or errors in the NLP analysis of the corpus. When only taking into consideration the errors directly related with the present work, the precision of the generation module was 86.7% in this evaluation.

In Table 5.3, the precision measure for each rule is presented. Precision on rule 4 (52.5%), with sentences that have indefinite subjects (pronouns or determiners), is 16.3% lower than the average system precision.

For each incorrect exercise, the error causes were annotated by the expert. The following causes were found:

pp-attach PP-attachment problem, which denounces a problem in the complement delimitation (described in section 4.4);

vdic-subj Incorrect identification of the inverted subject in a *verbum dicendi* construction, i.e. verbs that express or report speech, or introduce a quotation, e.g. “Não faça isso!”, disse o Pedro. (“Don’t do that”, said Peter).

Table 5.3: Evaluation precision for each rule.

<i>Rule</i>	<i>Precision</i>	<i>Correct Exercises #</i>	<i>Total Exercises #</i>
1 (main clauses)	72.5%	29	40
2 (verbal chains)	72.5%	29	40
3 (negation)	70%	28	40
4 (indefinite subjects)	52.5%	21	40
5 (adverbs)	65%	26	40
6 (subclauses)	80%	32	40
Total	68.8%	165	240

clit-pos Wrong clitic positioning among the answers;

pos-tag Incorrect POS tagging, for example, a preposition, e.g. *a*, incorrectly tagged as a definite article.

morph-v Incorrect attachment of the pronoun to the verb, resulting in incorrect enclisis instead of mesocclisis.

other Other causes, such as fixed expressions marked as direct complement (*valer a pena*), or corpus errors (e.g. non-grammatical sentences).

Table 5.4 presents the number of occurrences of each error class. Note that each exercise can have more than one error, therefore the total in this table is higher than the number of exercises.

Some causes are related to errors or shortcomings in the STRING processing chain analysis (the PP-attachment problem, the incorrect parsing of the subject of the *verba dicendi*, and POS tagging errors). Others are directly related to the present work (clitic positioning and mesocclisis).

The PP-attachment problem (described in section 4.4) was the most prevalent. The linguistic information in the corpus analysis is not sufficient to solve this problem. One way to avoid it would be to filter all sentences with complements that are followed by a prepositional phrase. However, this filter could remove too many sentences, and while this error makes the complement selection fail, it does not compromise the correct choice of case, form and position.

The morph-v error occurred because the future and conditional tenses were not being filtered in auxiliary verbs (only in main verbs), and the pronominalization verb termination for those tenses (mesocclisis) is not yet implemented. This filter is now correctly performed.

Some underlying causes for errors were identified in more detail, and some corrections to be performed in future work are presented in section 6.2.

Table 5.4: Incorrect exercises by error class.

<i>Error</i>	<i>Incorrect Exercises #</i>	<i>Incorrect Exercises %</i>	<i>Total Exercises %</i>
pp-attach	33	44.0%	13.8%
other	23	30.7%	9.6%
clit-pos	9	12.0%	3.8%
vdic-subj	12	16%	5%
morph-v	4	5.3%	1.7%
pos-tag	3	4.0%	1.3%

5.3 Crowd-sourced Test Results

5.3.1 Native Speakers Results

The results presented in this section were obtained from 114 native speakers (NS), with an average age of 31.5, ranging from 18 to 61 years old.

In Table 5.5, the number of incorrect answers by clitic positioning rule is shown. While these rules only directly affect clitic positioning, they result in sentences with different degrees of complexity, which might affect other factors in the answers. Main clauses have the fewest incorrect answers, being the simpler sentences. While verbal chains have the most complex structures and rules, they do not exhibit a higher error percentage than average. The highest number of incorrect answers happens with sentences that have indefinite subjects (pronouns or determiners). These sentences also happen to be the ones with more exercises deemed erroneous by the users (as seen in Table 5.6, possibly explaining the incorrect answers in some cases).

The number of exercises with errors as signaled by the users can be seen in Table 5.6. However, it should be taken into consideration that the reasons for each error can be very distinct. As mentioned in

Table 5.5: Incorrect answers by rule for NS.

<i>Rule</i>	<i>Incorrect Answers %</i>	<i>Incorrect Answers #</i>	<i>Total Answers</i>
1 (main clauses)	10.9%	12	110
2 (verbal chains)	20.8%	22	106
3 (negation)	19.8%	20	101
4 (indefinite subjects)	50.5%	50	99
5 (adverbs)	28.1%	27	96
6 (subclauses)	25.0%	23	92
Total	25.5%	154	604

Table 5.6: Number of exercises deemed erroneous by the NS users.

<i>Rule</i>	<i>Reported Errors #</i>	<i>Expert Agreement #</i>	<i>Expert Agreement %</i>
1 (main clauses)	18	10	55.6%
2 (verbal chains)	10	4	40%
3 (negation)	13	8	61.5%
4 (indefinite subjects)	28	24	85.7%
5 (adverbs)	14	11	78.6%
6 (subclauses)	13	10	76.9%
Total	96	67	69.8%

section 5.3.3, some users complained of lack of context in the exercise sentences, and such phenomenon is difficult to solve with the presented approach, but doesn't necessarily correlate with pronominalization errors. Other than that, the agreement between the NS and the expert was above 50% for all rules except with verbal chains; one possible explanation for this result is that clitic positioning within verbal chains often has more than one correct answer, but only the most general is presented among the multiple choices; in some cases, that answer, while it might be right, may not sound as canonical to some NS as other correct answers that were not effectively shown to the student.

In Figure 5.3, the distribution of incorrect answers by distractor type is shown. The following distractor types are presented:

pos Answers with the pronoun in the wrong position.

case Answers with the pronoun in the wrong case.

form Answers with the pronoun in the wrong accusative form.

pos+case Answers with the pronoun in the wrong position and with the wrong case.

Most errors occur with position distractors, as expected, since this is the linguistic phenomenon exhibits the most complex set of restrictions. However, though the choice of the pronoun case can be considered to constitute a simpler set of restrictions (agreement with the complement case), the case distractors are the second most common error found. A more in-depth linguistic investigation should be conducted to understand this phenomenon.

5.3.1.1 NS Questionnaire Results

The graphs in Figures 5.4 to 5.9 show the questionnaire results for each statement.

The majority users agreed that the system was easy to use, and that they quickly understood the objective of the exercises.

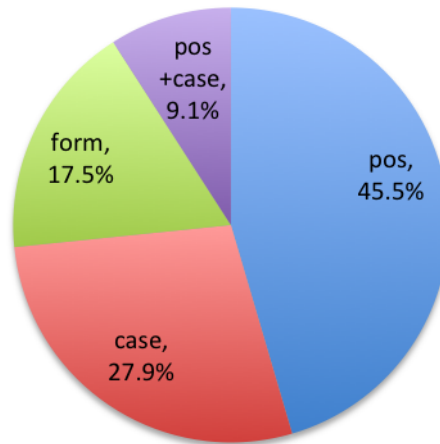


Figure 5.3: Distribution of incorrect answers by distractor type.

The statement about exercise difficulty had less agreement between evaluation subjects. 38% thought the difficulty was acceptable (not too easy or too difficult). 37% disagreed or strongly disagreed that the exercises were too easy, noting that they may be difficult, even for native speakers. On the other side, 26% agreed or strongly agreed that the exercises were too easy.

The majority of the users also agreed that the feedback was sufficient explanation for the answers.

More notably, 71% agreed or strongly agreed that the system is useful and they learned something by using it. This percentage is notable taking into consideration the users were native speakers.

As for the global appreciation of the system, the vast majority (85%) were somewhat or very satisfied.

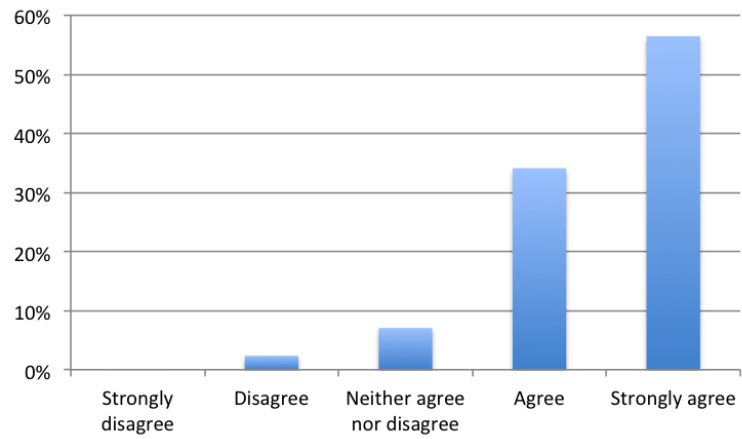


Figure 5.4: Results for the statement “The system is easy to use” for NS.

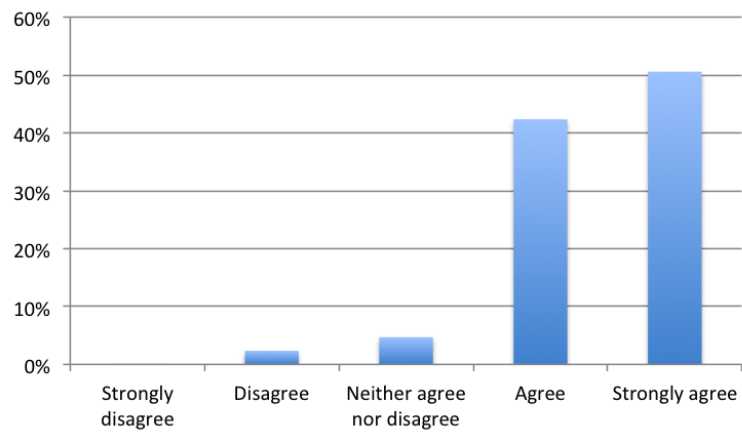


Figure 5.5: Results for the statement “I understood the objective quickly” for NS.

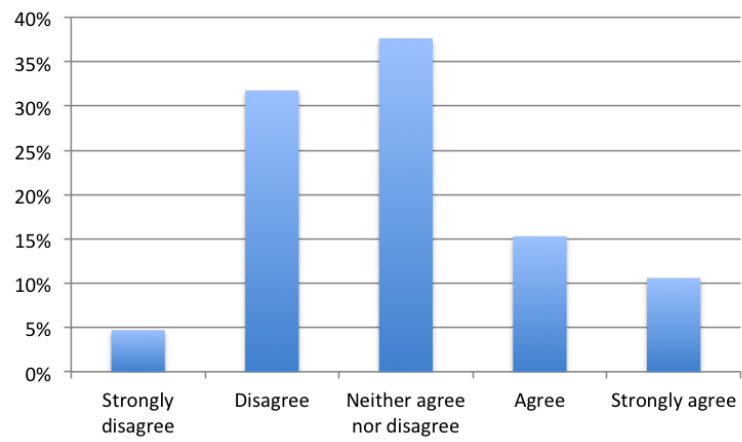


Figure 5.6: Results for the statement “The exercises are too easy.” for NS.

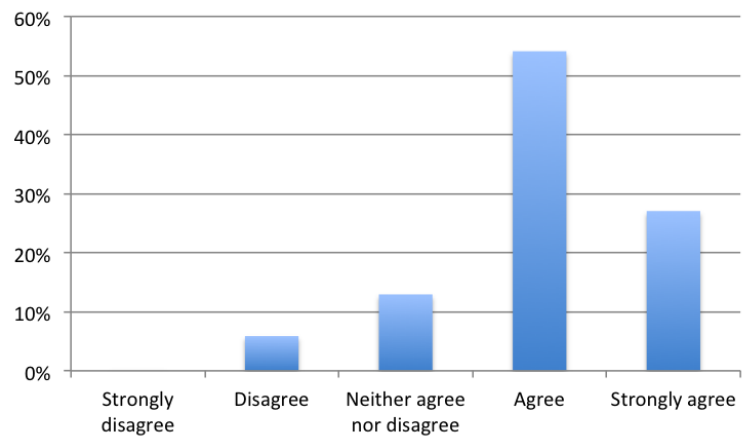


Figure 5.7: Results for the statement “The presented feedback is sufficient.” for NS.

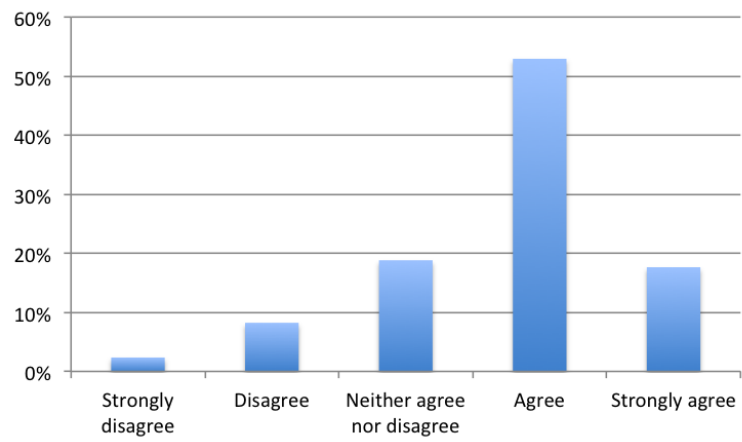


Figure 5.8: Results for the statement "The system is useful: I learned something by using it" for NS.

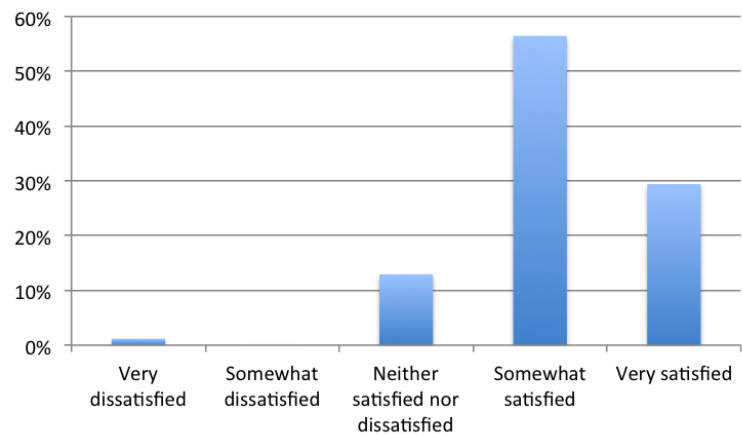


Figure 5.9: Results for the statement "Global system appreciation" for NS.

5.3.2 Non-Native Speakers Results

The results presented in this section were obtained from 19 non-native speakers (NNS), with an average age of 31.8, ranging from 20 to 60 years old. Three users studied Portuguese as a second language since childhood, and 9 studied Portuguese for more than 10 years. The number of years of Portuguese practice ranged from 1 to 33 years. Mother languages were English, Spanish, Italian, Russian and French.

In Table 5.7, the number of incorrect answers by clitic positioning rule is shown. The incorrect answers appear uniformly distributed among the positioning rules, with an average of 29%. Clauses with adverbs had the fewest incorrect answers. As seen with native speakers, sentences that have indefinite subject (pronouns or determiners) have a higher than average error rate. The highest number of incorrect answers happens in subordinate clauses.

NNS users did not signal any exercise as having errors, potentially because they were not confident in their knowledge to do so.

In Figure 5.10, the distribution of incorrect answers by distractor type for NNS is shown. The distractor type combining position and case errors were the most common, showing that this combination is more challenging for NNS than for NS (51.9% vs 9.1%). The form distractor error rate was similar for NNS and NS (22.2% vs 17.5%).

5.3.2.1 NNS Questionnaire Results

The graphs in Figures 5.11 to 5.16 show the questionnaire results for each statement.

All NNS users agreed or strongly agreed that the system was easy to use, and most agreed they quickly understood the objective of the exercises.

As with NS, the statement about exercise difficulty had less agreement between evaluation subjects. 13% thought the difficulty was acceptable (not too easy or too difficult); 40% disagreed that the exercises

Table 5.7: Incorrect answers by rule for NNS.

<i>Rule</i>	<i>Incorrect Answers %</i>	<i>Incorrect Answers #</i>	<i>Total Answers</i>
1 (main clauses)	29.4%	5	17
2 (verbal chains)	26.7%	4	15
3 (negation)	26.7%	4	15
4 (indefinite subjects)	33.3%	5	15
5 (adverbs)	20.0%	3	15
6 (subclauses)	37.5%	6	16
Total	29.0%	27	93

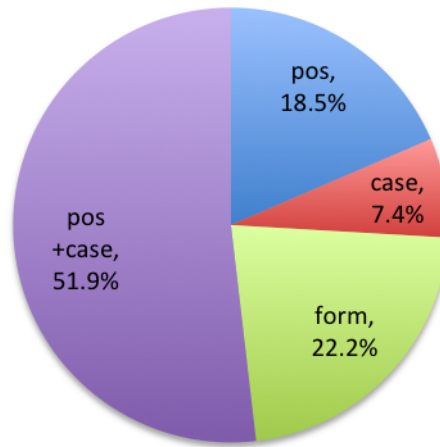


Figure 5.10: Distribution of incorrect answers by distractor type for NNS.

were too easy, noting that they may be difficult. On the other side, 47% agreed that the exercises were too easy. This split could be explained by the NNS age and proficiency distribution.

Almost all the NNS users (87%) also agreed or strongly agreed that the feedback was sufficient explanation for the answers. None disagreed, compared to the 6% NS that found the feedback could be more detailed, or with more examples as seen in the comments.

80% of the NNS agreed or strongly agreed that the system is useful and they learned something by using it, a 9% increase from NS. Every NNS considered to have learned something, compared to 10% of NS that did not considered the system useful.

As for the global appreciation of the system, the same percentage (85%) were somewhat or very satisfied.

5.3.3 Questionnaire Comments

In the free-form text comments at the end of the questionnaire, several problems were raised and suggestions were made:

Lack of context The most frequent comments were on the lack of context of many sentences. Since they are taken from larger texts in the corpus, some sentences do not make much sense taken out of their context. In some cases, after the pronominalization the sentence becomes unintelligible for the lack of antecedent that had occurred in a previous sentence, not presented in the exercise.

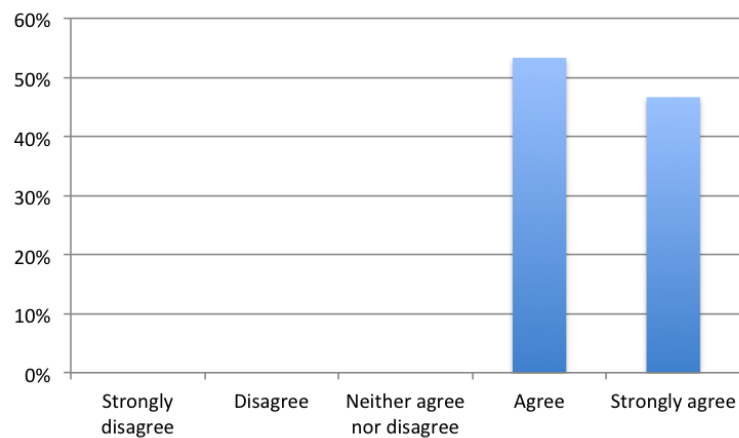


Figure 5.11: Results for the statement “The system is easy to use” for NNS.

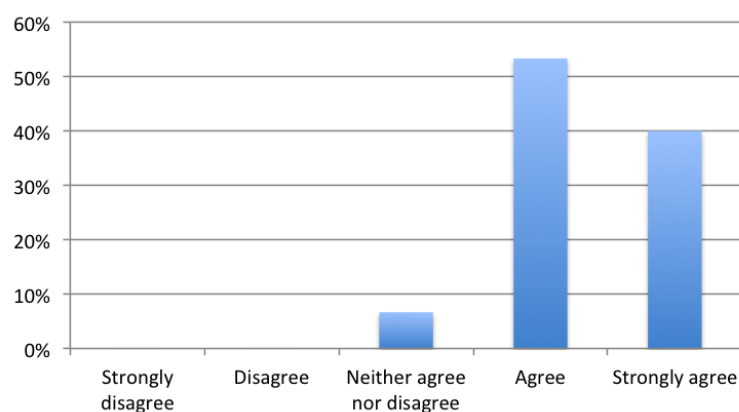


Figure 5.12: Results for the statement “I understood the objective quickly” for NNS.

A clarification of the reasons of this phenomenon before the exercises begin was suggested, to minimize confusion on the part of the learners, since they may think it’s a generation error. This problem could be minimized by developing more sentence filters to eliminate some sentences and complements that need a context to be understood. For example, eliminating sentences that begin with a conjunction, e.g. *e* (and) or *mas* (but). However, ultimately this exercise does not aim to teach the students in what circumstances they should pronominalize a constituent, only how to do a correct pronominalization. In order to add context to the exercise sentences, since the same complement is not usually used twice in consecutive sentences, anaphora resolution would have to be used to generate the question sentence (so the original sentence would be the correct answer). This alternative approach is complex and has other problems, as explained in sections 2.4 and 4.3.

Too complex explanations Some users complained that the feedback explanations were too complex

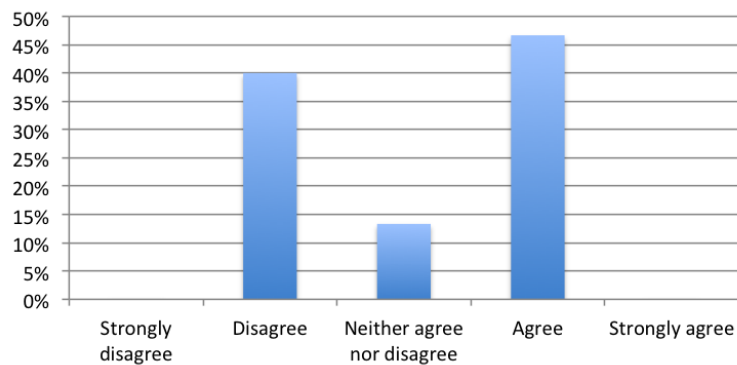


Figure 5.13: Results for the statement “The exercises are too easy” for NNS.

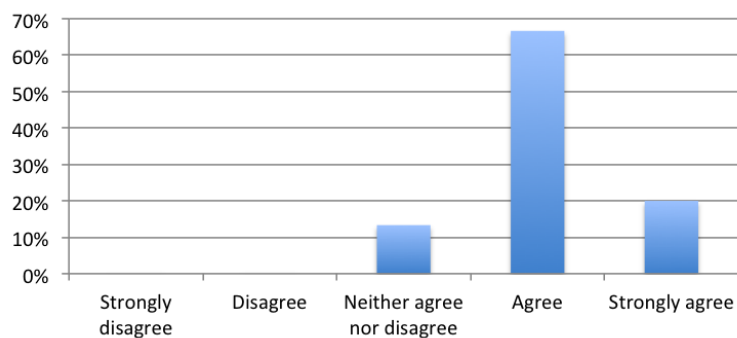


Figure 5.14: Results for the statement “The presented feedback is sufficient” for NNS.

and used technical language. In order to be more accessible to a wide range of language learners, the feedback language could be simplified, with links to more complex explanations. Alternatively, the texts could be adapted to the student’s language proficiency and number of years of contact with the language. Other comments suggested the addition of more examples. On the other hand, most comments praised the detailed feedback and ability to learn new aspects, even when the answers were correct.

Feedback for the wrong answer Apart from the explanations about the correct answer, one comment suggested that the system explained what was wrong in the selected incorrect answer. This possibility was considered, and partially implemented, in the interface development, but discarded in order to maintain simplicity and avoid confusing the students. It would be easy to finish implementation in future developments.

Incorrect complement selection Several comments noted the incorrect complement selection, or consequent errors in the pronominalization, where the sentence stops making semantic sense or be-

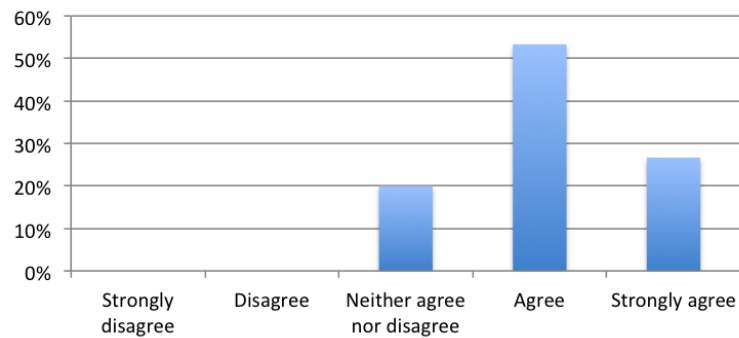


Figure 5.15: Results for the statement “The system is useful: I learned something by using it” for NNS.

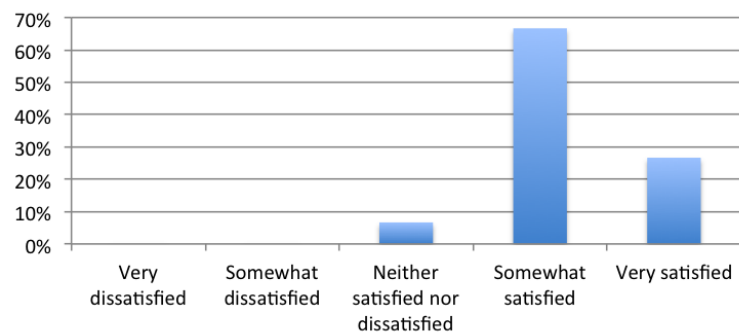


Figure 5.16: Results for the statement “Global system appreciation” for NNS.

comes agrammatical. This is due to the PP-attachment problem previously noted in section 5.2 and described in section 4.4.

Webpage design Some comments suggested better web design to improve appeal and usability, e.g. using different fonts and colors.

Conclusion and Future Work

6.1 Final Remarks

In an increasingly competitive and dynamic world, it is essential that innovative approaches are developed in the education area and in language education in particular. REAP.PT is a pioneer project in an emergent interdisciplinary field, and presents important challenges in the integration of language teaching exercises to its personalized and dynamic model that makes it appealing to students and teachers.

We believe that the work developed in this dissertation is a valuable new asset for the creation of new syntactic exercises for the European Portuguese language. Several good practices to be adopted in the future were described. The general architecture of the REAP.PT syntactic module, specifically the choice of technologies is expected to make a relevant step forward in order to ease the development effort and factorize the common code between modules and future exercises. The pioneer feedback system with detailed and automatically generated explanations for each answer is also believed to be an asset for future exercises, and was praised by users, improving the quality of the learning experience and its efficiency.

Some pitfalls were also uncovered during the development, such as the unapparent complexity of some aspects of syntactic exercise generation, that were only unfolded as the development progressed (e.g. the pronoun positioning rules). Heavy reliance on correctness and completeness of the NLP analysis of the text is also a factor to be taken into account (e.g. the PP-attachment problem or the need for several sentence filters). Therefore, the analysis of the exercise generation approach and NLP analysis information needs is very important in the success of its development, and should be performed thoroughly in the initial phases.

This work contributed to the improvement of the STRING processing chain, by identifying shortcomings, such as focus adverbs (as seen in Listing 6.1¹), and areas of future work, including some whose importance was not evident before their practical application, namely the importance of the identification of the subject in *verbum dicendi* constructions.

Listing 6.1: Pronominalization exercise example.

```
1 <LUNIT start="68840" end="68902">  
2   <original rule="3" comp="os videoclubes"  
3     file="/corpora/publico/20121004/Parte10/Parte10ael.out" verb="atinge">
```

```

4      E a crise não atinge só [[ os videoclubes ]], mas também as editoras.
5      </original>
6      <answer>
7          <response accusative="true" position_after="false">
8              E a crise não {{os}} atinge só, mas também as editoras.
9          </response>
10     </answer>
11     <distractors>
12         <response accusative="false" position_after="true">
13             E a crise não atinge—{{lhes}} só, mas também as editoras.
14         </response>
15         <response accusative="false" position_after="false">
16             E a crise não {{lhes}} atinge só, mas também as editoras.
17         </response>
18         <response accusative="true" position_after="true">
19             E a crise não atinge—{{os}} só, mas também as editoras.
20         </response>
21     </distractors>
22 </LUNIT>

```

6.2 Future Work

For some errors detected during the evaluation, the cause was identified and we propose corrections for future work.

When the direct complement is in the infinitive, introduced by the proposition *a*, the clitic cannot be positioned between the preposition and the verb, e.g.: *O Pedro obrigou a Ana a ler o livro*.

**O Pedro obrigou a Ana a o ler*.

If there are other elements such as adverbs between the preposition and the main verb, the clitic can be introduced before the verb, but not immediately after the preposition, e.g.: *O Pedro obrigou a Ana a imediatamente ler o livro*.

* *O Pedro obrigou a Ana a o imediatamente ler*.

O Pedro obrigou a Ana a imediatamente o ler.

Also in infinitives introduced by prepositions, the clitic can be in the pre-verbal position, as seen in this example from the corpus, where the following positioning was incorrectly marked as a distractor:

Acusa Hollywood de retratar sempre os homossexuais e lésbicas como psicóticos ou assassinos.

Acusa Hollywood de os retratar sempre como psicóticos ou assassinos.

Other suggestions of future work are the following:

¹In this case, the use of a focus adverb (*só* “only”) could have complicated matters. However, because of its complex syntax, this adverb has not been given the *focus adverb* feature yet, pending on further development of the STRING system. Therefore, *só* is left out from the *NP* that is targeted by the pronominalization.

Generate exercises from other corpora Texts of different genres other than news, such as literary texts or cuisine recipes could be used to generate exercises with more variety. In particular, it may be necessary for NNS to use simpler texts, which imply textual complexity filters, in order to adapt them to the learners proficiency and to improve the NLP results/analysis;

Post-editing interface The teachers should be able to manually evaluate, discard or modify the generated exercises. Both the sentence and answers should be modifiable, and the reasons for the modifications could be added so the system can be improved;

Future-indicative and conditional tenses The mesoclis (where the clitic is placed between the thematic vowel and the verb tense endings) has yet to be implemented;

Simultaneous case pronominalization The current exercise was built so as to generate only one pronoun case at a time. If there are both direct and indirect complements for the same verb, they could be pronominalized at the same time (e.g. *O Pedro leu o livro ao João* = *O Pedro leu-lho*, “Peter read **the book** to John = Peter read **it-him**.”). The resulting contractions could be taught to the students without relying on their occurrence in the corpus (when a sentence already has one of the complements pronominalized). This can be done in the current architecture by re-analyzing the generated pronominalization and pronominalizing the other complement, but could be made more efficient. However, there were not enough indirect complement dependencies generated by the current version of the STRING processing chain to justify this feature at present;

Interface Caching Introduce PHP bytecode caching to improve performance. This was not needed in the evaluation as performance was good, but with more users, the distractor form and feedback page generation could become a bottleneck. Caching would be an easy and effective way to solve this potential problem. PHP APC cache² or eAccelerator³ are suggested.

²<http://php.net/manual/en/book.apc.php> (last visited in October 2012)

³<https://github.com/eaccelerator/eaccelerator> (last visited in October 2012)

Bibliografia

- Aït-Mokhtar, S., Chanod, J.-P., & Roux, C. (2002, June). Robustness beyond shallowness: incremental deep parsing. *Nat. Lang. Eng.*, 8(3), 121–144.
- Aldabe, I. (2011). *Automatic Exercise Generation Based on Corpora and Natural Language Processing Techniques*. Unpublished doctoral dissertation, Euskal Herriko Unibertsitatea (University of the Basque Country), San Sebastian, Basque Country.
- Aldabe, I., Lacalle, M. L. de, Maritxalar, M., & Martinez, E. (2007). The Question Model inside ArikIturri. In J. M. Spector et al. (Eds.), *Proceedings of the 7th IEEE International Conference on Advanced Learning Technologies, ICALT 2007, July 18-20 2007, Niigata, Japan* (p. 758-759). IEEE Computer Society.
- Amaral, L., & Meurers, D. (2011). On Using Intelligent Computer-Assisted Language Learning in Real-Life Foreign Language Teaching and Learning. *ReCALL*, 23(1), 4–24.
- Baptista, J. (2012, July). *Positioning of Clitic Pronouns in European Portuguese (Working Paper)*.
- Baptista, J., Costa, N., Guerra, J., Zampieri, M., Cabral, M., & Mamede, N. J. (2010). P-AWL: Academic Word List for Portuguese. In T. A. S. Pardo, A. Branco, A. Klautau, R. Vieira, & V. L. S. de Lima (Eds.), *Computational Processing of the Portuguese Language, 9th International Conference, PROPOR 2010, Porto Alegre, RS, Brazil, April 27-30, 2010. Proceedings* (Vol. 6001, p. 120-123). Springer.
- Benzaken, V., Castagna, G., & Frisch, A. (2003, August). CDuce: an XML-centric general-purpose language. *SIGPLAN Not.*, 38(9), 51–63.
- Chamberlin, D. (2003). XQuery: a query language for XML. In *Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data* (pp. 682–682). New York, NY, USA: ACM.
- Chen, C.-Y., Liou, H.-C., & Chang, J. S. (2006). FAST: an automatic generation system for grammar tests. In *Proceedings of the COLING/ACL on Interactive Presentation Sessions* (pp. 1–4). Stroudsburg, PA, USA: Association for Computational Linguistics.
- Correia, R. (2010). *Automatic Question Generation for REAP.PT Tutoring System*. Unpublished master's thesis, Instituto Superior Técnico - Universidade Técnica de Lisboa, Portugal.

- Correia, R., Baptista, J., Eskenazi, M., & Mamede, N. J. (2012). Automatic Generation of Cloze Question Stems. In H. de Medeiros Caseli, A. Villavicencio, A. J. S. Teixeira, & F. Perdigão (Eds.), *Computational Processing of the Portuguese Language - 10th International Conference, PROPOR 2012, Coimbra, Portugal, April 17-20, 2012. Proceedings* (Vol. 7243, p. 168-178). Springer.
- Costa, F., & Mendonça, L. (2011). *Diálogos Caderno de Actividades*. Porto Editora.
- Emir, B. (2003). *Extending pattern matching with regular tree expressions for XML processing in Scala*. Unpublished master's thesis, RWTH Aachen.
- Gapeyev, V., Levin, M., Pierce, B., & Schmitt, A. (2005). XML goes native: Run-time representations for Xtatic. In *Compiler Construction* (pp. 138–138).
- Hosoya, H., & Pierce, B. C. (2003, May). XDuce: A statically typed XML processing language. *ACM Trans. Internet Technol.*, 3(2), 117–148.
- Mamede, N., Baptista, J., & Hagège, C. (2011, May). *Nomenclature of Chunks and Dependencies in Portuguese XIP Grammar 3.1* (Tech. Rep.). Lisbon: L2F/INESC-ID.
- Mamede, N. J., Baptista, J., Diniz, C., & Cabarrão, V. (2012, April). *STRING: An Hybrid Statistical and Rule-Based Natural Language Processing Chain for Portuguese*. <http://www.propor2012.org/demos/DemoSTRING.pdf>.
- Marques, C. (2011). *Syntactic REAP.PT*. Unpublished master's thesis, Instituto Superior Técnico - Universidade Técnica de Lisboa, Portugal.
- Marujo, L. (2009). *REAP em Português*. Unpublished master's thesis, Instituto Superior Técnico - Universidade Técnica de Lisboa, Portugal.
- Mendes, A. C., Curto, S., & Coheur, L. (2011). Bootstrapping multiple-choice tests with THE-MENTOR. In *Proceedings of the 12th international conference on Computational linguistics and intelligent text processing - Volume Part I* (pp. 451–462). Berlin, Heidelberg: Springer-Verlag.
- Meurers, D., Ziai, R., Amaral, L., Boyd, A., Dimitrov, A., Metcalf, V., et al. (2010, June). Enhancing Authentic Web Pages for Language Learners. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 10–18). Stroudsburg, PA, USA: Association for Computational Linguistics.
- Móia, T., & Peres, J. A. (2003). *Áreas Críticas da Língua Portuguesa*. Lisboa: Editorial Caminho.
- Moreira, J. E., Michael, M. M., Silva, D. D., Shiloach, D., Dube, P., & Zhang, L. (2007). Scalability of the Nutch search engine. In B. J. Smith (Ed.), *Proceedings of the 21th Annual International Conference on Supercomputing, ICS 2007, Seattle, Washington, USA, June 17-21, 2007* (p. 3-12). ACM.

- Nobre, N. (2011). *Anaphora Resolution*. Unpublished master's thesis, Instituto Superior Técnico - Universidade Técnica de Lisboa.
- Paulo, S., Oliveira, L. C., Mendes, C., Figueira, L., Cassaca, R., Viana, C., et al. (2008). DIXI - A Generic Text-to-Speech System for European Portuguese. In A. J. S. Teixeira, V. L. S. de Lima, L. C. de Oliveira, & P. Quaresma (Eds.), *Computational Processing of the Portuguese Language, 8th International Conference, PROPOR 2008, Aveiro, Portugal, September 8-10, 2008, Proceedings* (Vol. 5190, p. 91-100). Springer.
- Pellegrini, T., Correia, R., Trancoso, I., Baptista, J., & Mamede, N. J. (2011). Automatic Generation of Listening Comprehension Learning Material in European Portuguese. In *INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27-31, 2011* (p. 1629-1632). ISCA.
- Santos, D., & Rocha, P. (2001). Evaluating CETEMPUBLICO, a Free Resource for Portuguese. In *Association for Computational Linguistics, 39th Annual Meeting and 10th Conference of the European Chapter, Proceedings of the Conference, July 9-11, 2001, Toulouse, France* (p. 442-449). Morgan Kaufmann Publishers.
- VVAA. (2010). *Português dez - Caderno de apoio ao aluno*. Lisboa: Lisboa Editores.

I Appendices

A

XQuery Rule Example

Listing A.1: XQuery Rule 1 - Simplest case of affirmative main clauses without verbal chains.

```
1 import module namespace pronLib = "http://call.l2f.inesc-id.pt/reap.  
  public/pronominalizationLib" at "libs/pronominalizationLib.xquery";  
2  
3 declare variable $case_accusative external := true();  
4  
5 declare variable $rulename := "1";  
6  
7 let $lunits := /*/LUNIT  
8  
9   for $lunit in $lunits  
10  
11     let $lunit := pronLib:basicFilters($lunit, $case_accusative)  
12     return if (not($lunit)) then () else  
13  
14  
15     let $const_name := if($case_accusative) then "CDIR" else "CINDIR"  
16     let $consts_dep := pronLib:selectConstituents($lunit, $const_name)  
17     return if (not($consts_dep)) then () else  
18  
19     (: generate exercises from all the constituents on the sentence :)  
20     for $const_dep in $consts_dep  
21  
22       let $verb_node := pronLib:getVerb($lunit, $const_dep/PARAMETER[1]/@num, false())  
23       let $verb_num := ($const_dep/PARAMETER[1]/@num, $verb_node/@num)  
24       return if (not($verb_node)) then () else  
25  
26       (: all complements of the same verb :)  
27       let $consts := $lunit/DEPENDENCY[@name=$const_name and  
28         PARAMETER[1]/@num=$const_dep/PARAMETER[1]/@num]  
29  
30       (: verb modifiers :)  
31       let $verb_mods := $lunit/DEPENDENCY[@name="MOD" and PARAMETER[1]/@num=$verb_num]  
32  
33       let $lunit := $lunit[  
34  
35         (: positive clauses :)  
36         not(pronLib:negativeVerbMods($verb_mods))  
37  
38         (: case 4 subject modifiers :)  
39         and not(pronLib:subjMods($lunit, $verb_num))  
40  
41         (: no adverbs modifying verb :)  
42         and not(pronLib:adverbMods($lunit, $verb_mods))  
43
```

```

44     (: no verbal chains :)
45     and not(pronLib:auxVerbs($lunit, $verb_num))
46
47 ]
48
49 return if(not($lunit)) then () else
50
51     (:selects constituent:)
52     let $const := pronLib:calcConst($lunit, $consts)
53     return if (not($const instance of map(*))) then () else
54
55     (: try to filter untagged subordinated clauses ("que" before the constituent and
56        after the previous punct) :)
57     let $prec_punct := $const('nodes')[1]/preceding::NODE[@tag="PUNCT"][1]
58     let $que := $lunit/NODE//NODE[not(/child::NODE) and (. >> $prec_punct or not($prec_punct))
59         and . << $const('nodes')[1] and TOKEN/READING/@lemma="que"]
60     return if ($que) then () else
61
62     let $pronominalized :=
63     pronLib:pronominalize($lunit, $const, $verb_node, $case_accusative, true())
64
65     return if ($pronominalized)
66     then
67         let $verb_str := replace(($verb_node/TOKEN/text())[1], '\s', '')
68         let $distractors :=
69             (pronLib:pronominalize($lunit, $const, $verb_node, not($case_accusative)
70                 , true()),
71              pronLib:pronominalize($lunit, $const, $verb_node, not($case_accusative)
72                 , false()),
73              pronLib:pronominalize($lunit, $const, $verb_node, $case_accusative, false
74                 ()))
75
76         return pronLib:print-exercise(
77             $lunit, $const, $pronominalized, $distractors, $rulename, ("verb"), (
78                 $verb_str))
79
80     else ()

```

Clitic Positioning within verbal chains: Empirical Study

				Aux Prep V-C	Aux Prep C V	Aux-C Prep V	C Aux Prep V	Neg Aux Prep V-C	Neg Aux Prep C V	Neg Aux-C Prep V	Neg C Aux Prep V	SC Aux Prep V-C	SC Aux Prep C V	SC Aux-C Prep V	SC C Aux Prep V	Tot	%
Aux	Prep	Verb Type	Aux Type	1	2	3	12	4	5	6	7	8	9	10	11		
acabar	a	VINF	VASP	4								2	4			10	0.1%
acabar	de	VINF	VASP	8	59				1			5	30		12	115	0.8%
acabar	por	VINF	VASP	358	296		4	1	4			109	138		9	919	6.8%
andar	a	VINF	VASP	53	4		4	5	4		4	30	72		53	229	1.7%
cessar	de	VINF	VASP						10				6			16	0.1%
chegar	a	VINF	VASP	101	6	3	10	50	9		31	60	8		48	326	2.4%
começar	por	VINF	VASP	37	28							16	8			89	0.7%
começar	a	VINF	VASP	256	35	1	7	2	1		5	201	306	1	66	881	6.5%
continuar	a	VINF	VASP	385	10	2	4	1	1		2	340	49		69	863	6.4%
correr	a	VINF	VASP	10					2			9				21	0.2%
cuidar	de	VINF	VMOD		6		1				4	1	4		1	17	0.1%
deixar	de	VINF	VASP	8	273	9	7	10	208		8	35	510	4	27	1099	8.1%
desatar	a	VINF	VASP	5								1	5			11	0.1%
estar	para	VINF	VASP	1	11			2	14		1	3	48		1	81	0.6%
estar	a	VINF	VASP	572	74	38	22	45	80		41	325	1362	1	510	3070	22.7%
ficar	de	VINF	VMOD	1	5	4						1	5			16	0.1%
ficar	a	VINF	VASP	56	9	1	4	3	5		4	22	50	1	11	166	1.2%
haver	que	VINF	VMOD	226	43			14	4			10	10			307	2.3%
haver	de	VINF	VMOD	63	34		8	8	4		8	45	45		177	392	2.9%
hesitar	em	VINF	VMOD	4				80	16			35	5			140	1.0%
ir	a	VINF	VTEMP	1	9						1	2	4		1	18	0.1%
parar	de	VINF	VASP	2	2			2	27			2	19			54	0.4%
passar	a	VINF	VASP	166	5	29	1	4	3			101	74	2	23	408	3.0%
ser	de	VINF	VMOD	11	14			11	2		4	5	2		4	53	0.4%
tender	a	VINF	VMOD	45	1			3				36			2	87	0.6%
ter	que	VINF	VMOD	67	583		3	1	63		2	27	238	1	29	1014	7.5%
ter	de	VINF	VMOD	168	700		7	8	35		5	91	397		60	1471	10.9%
tornar	a	VINF	VASP	23	1	8		1		1	5	11			22	72	0.5%
tratar	de	VINF	VMOD	20	55			1	7			5	38		2	128	0.9%
vir	a	VINF	VASP	98	52		2	1	2		2	241	120		220	738	5.5%
voltar	a	VINF	VASP	316		3	8	12			10	258	10		106	723	5.3%
			tot	3065	2315	98	92	265	502	1	137	2029	3567	10	1453	13534	
			%	22.6%	17.1%	0.7%	0.7%	2.0%	3.7%	0.0%	1.0%	15.0%	26.4%	0.1%	10.7%		

Table B.1: Clitic positioning counts on auxiliary verbs with linking prepositions.

			Aux Prep V-C	Aux Prep C V	Aux-C Prep V	C Aux Prep V	Neg Aux Prep V-C	Neg Aux Prep C V	Neg Aux-C Prep V	Neg C Aux Prep V	SC Aux Prep V-C	SC Aux Prep C V	SC Aux-C Prep V	SC C Aux Prep V		
Aux	Verb Type	Aux Type	1	2	3	12	4	5	6	7	8	9	10	11	Tot	%
conseguir	VINF	VMOD	450		5	41	240	1		283	469			402	1891	6.8%
costumar	VINF	VASP	45		1	1	3			3	21	1		24	99	0.4%
dever	VINF	VMOD	845	3	29	42	124			198	281	1	2	660	2185	7.9%
estar	VGER	VASP	5	4	1	1				1	1	1		13	27	0.1%
haver	VPP	VTEMP			30	10				19		1	2	266	328	1.2%
ir	VINF	VTEMP	1888	5	151	103	140		1	315	671	1	4	1915	5194	18.7%
ir	VGER	VASP	22		172	21					6		3	280	504	1.8%
ousar	VINF	VMOD	24				10	1		1	55	1		5	97	0.3%
poder	VINF	VMOD	1939	3	48	294	571	25		1014	1260	14	3	2336	7507	27.0%
procurar	VINF	VMOD	240		2	2	8	4			156	1		58	471	1.7%
tencionar	VINF	VMOD	63				16			5	49			11	144	0.5%
tentar	VINF	VMOD	845		6	18	45	7		6	940	1	4	455	2327	8.4%
ter	VPP	VTEMP			917	278		5	8	339		7	675	4507	6736	24.3%
ter	VINF	VMOD		2				7				5			14	0.1%
terminar	VGER	VASP	7		3						2				12	0.0%
vir	VINF	VASP	23		28	1	1			1	35	1	5	15	110	0.4%
vir	VGER	VASP	1		6	3				2				96	108	0.4%
		tot	6397	17	1399	815	1158	50	9	2187	3946	35	698	11043	27754	
		%	23.0%	0.1%	5.0%	2.9%	4.2%	0.2%	0.0%	7.9%	14.2%	0.1%	2.5%	39.8%		

Table B.2: Clitic positioning counts on auxiliary verbs without linking prepositions.

C

Questionnaire

Pronominalização



Questionário

Obrigado pela sua participação.
Acertou de 6 perguntas.

A sua colaboração e opinião são muito importantes para nós.
Por favor, preencha o seguinte questionário:

1. O sistema é fácil de utilizar.

Discordo plenamente	Discordo	Nem concordo, nem discordo	Concordo	Concordo plenamente
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

2. Percebi rapidamente o objectivo.

Discordo plenamente	Discordo	Nem concordo, nem discordo	Concordo	Concordo plenamente
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

3. Os exercícios são demasiado fáceis.

Discordo plenamente	Discordo	Nem concordo, nem discordo	Concordo	Concordo plenamente
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

4. O feedback apresentado é suficiente.

Discordo plenamente	Discordo	Nem concordo, nem discordo	Concordo	Concordo plenamente
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

5. O sistema é útil: aprendi alguma coisa ao usá-lo.

Discordo plenamente	Discordo	Nem concordo, nem discordo	Concordo	Concordo plenamente
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

6. Apreciação global do sistema.

Muito insatisfeito	Insatisfeito	Nem satisfeito, nem insatisfeito	Satisfeito	Muito satisfeito
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Comentários Adicionais:

(Como podemos melhorar o sistema? De que gostou mais e/ou menos?)

Submeter

