Syntactic REAP.PT: Exercises on Clitic Pronouning*

Tiago Freitas[1], Jorge Baptista[2], and Nuno Mamede[1]

1 **IST – Instituto Superior Técnico**
L²F – Spoken Language Systems Laboratory – INESC ID Lisboa
Rua Alves Redol 9, 1000-029 Lisboa, Portugal
Tiago.Freitas@ist.utl.pt, Nuno.Mamede@ist.utl.pt
2 **Universidade do Algarve, FCHS/CECL**
Campus de Gambelas, 8005-139 Faro, Portugal jbaptis@ualg.pt

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. In this paper, we present a new improvement of the REAP.PT system, consisting in 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 often be uncertain about the 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 positive results demonstrated the validity of the present approach.

1998 ACM Subject Classification I.2.7 Natural Language Processing

Keywords and phrases Intelligent Computer Assisted Language Learning (ICALL), Portuguese, Syntactic Exercises, Automatic Exercise Generation, Clitic Pronouning

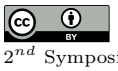
Digital Object Identifier 10.4230/OASIcs.SLATE.2013.271

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

* This work was supported by national funds through FCT – Fundação para a Ciência e a Tecnologia, under project PEst-OE/EEI/LA0021/2011 and under FCT project CMU-PT/HuMach/0053/2008.

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



2nd Symposium on Languages, Applications and Technologies (SLATE'13).
Editors: José Paulo Leal, Ricardo Rocha, Alberto Simões; pp. 271–285

OpenAccess Series in Informatics
OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

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 (REAder-specific Practice) project[2] is one of such systems, developed at CMU[3] by the LTI[4] 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.PT[6] 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 and question generation phases [13, 7]. More recently, a listening comprehension module was also developed [14]. The system was then extended to include syntax learning as well [12].

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. This 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 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.

2 Related Work

In this section, a brief review of the related work is made. Firstly, some automatic question generation systems for other languages are presented, then a succint description of current syntactic textbook exercises on pronominalization is done. Finally, section 3 describes the current architecture of the REAP.PT system, where this work is included.

2.1 ICALL Systems

There are not many ICALL systems that include automatic generation of exercises, and even less for syntactic exercises. FAST [6] (Free Assessment of Structural Tests) is an automatic

[2] `http://reap.cs.cmu.edu` (last visited in October 2012)
[3] Carnegie Mellon University - `http://www.cmu.edu` (last visited in October 2012)
[4] Language 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)

question generation system for grammar tests in the English language, using a method that involves representing the questions' characteristics as structural patterns (surface patterns made of POS tags), and applying those patterns in order to transform sentences into exercises (multiple-choice and error detection questions). Arikiturri [3, 2] is a modular and multilingual automatic question generation system. It is currently implemented for Basque and 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. It uses a question model to represent the exercises (as well as the information relating to their generation process). It also has a web-based post-editing environment.

2.2 Current Syntactic Exercises on Pronominalization

There are several pronominalization exercises in textbooks and on-line resources: given three forms of pronouns, choose the right one to replace the signalled constituent; correct and incorrect sentences, that must be classified according to clitic placement; given a small text with signalled pronouns, rewrite the text replacing the pronouns with their corresponding antecedents; given a declarative affirmative sentence with clitics, transform it to the corresponding negative sentence; and cloze questions, in which the student has to choose (multiple-choice) or fill in the correct pronoun to replace the signalled constituent.

The last type of exercise, cloze questions, is the easiest to automatically generate, since the exercises can be produced by manipulating the original sentence. Other exercises involve text generation, which is complex and less objective in their evaluation. [16] is a tool that helps teachers producing corpus-based cloze questions. However, no system was found for Portuguese that both automatically generates and corrects (cloze) questions, as it is aimed here.

3 REAP.PT Architecture

3.1 Old REAP.PT Architecture.

The initial REAP.PT architecture, focused on reading comprehension and vocabulary exercises, consists of several components. The Web Interface component is responsible for the user interaction with the system and information exchange between the database and the listening comprehension module. A listening comprehension module 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, text information, focus words and related vocabulary questions and distractors (or foils). The second database stores the lexical resources. A filter chain is used to select a subset of the corpus that fits within certain practical and pedagogical constraints [13]. The topic and readability classifiers run on the output of the filter chain and classify the texts according to topic and reading level [13]. The question generation module is responsible for the generation of vocabulary exercises given to the students after each text reading.

The work on the question generation module started in Correia [7], with a focus on vocabulary *cloze* (*fill-in-the-blank*) questions, and the study of the distractors, the wrong multiple-choice alternatives. The existing exercises include definition questions, synonym questions, hyperonym/hyponym questions, cloze questions about the text, and syntactic exercises.

The current syntactic exercises in REAP.PT [12] 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 [11], based on XIP [1], is used to extract relevant dependencies. Distractors are then generated using the L^2F VerbForms[7] word form generator for verbs, and a set of rule-based restrictions are applied to reduce ambiguity.

The 'Nominal Determinants' exercise aims to teach distributional constraints between a determinative noun and the noun it determines (e.g. *copo de leite* 'glass of milk'), and at the same time the relationship between collective names and common (e.g. *mata de cedros* 'wood of cedars'). A feedback system teaches the student the missed definitions, giving examples and images illustrative of the determinative nouns.

The architecture of the syntactic exercise generation can be seen in Figure 1. 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.

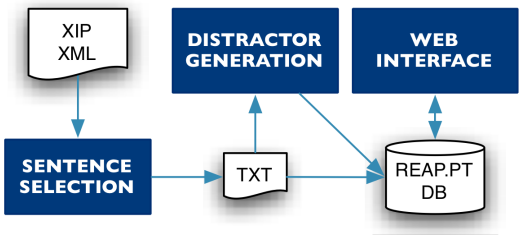


Figure 1 REAP.PT syntactic exercises architecture.

In the sentence selection phase, the XIP output is processed, and the syntactic features are analysed in order to select the stems that are to be used to generate the questions. This phase is performed using the Hadoop[8] 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.

3.2 New Exercises Generation

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 (fill-in-the-blank). 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).

The following challenges were considered: selection rules complexity, several different sentence types, and generation metadata for a feedback system. The intention behind this new architecture was not only to simplify the implementation of the proposed exercise, but

[7] `https://www.l2f.inesc-id.pt/wiki/index.php/VerbForms` (last visited in October 2012)
[8] `http://hadoop.apache.org` (last visited in October 2012)

also 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 2.

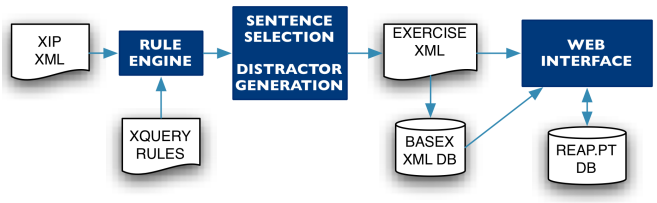


Figure 2 REAP.PT new syntactic exercises architecture.

In order to develop the exercises, the STRING [11] NPL processing chain is used to analyze the corpus sentences, which outputs the syntactic tree and dependencies in XML [10]. The need for a high-level XML processing language was identified, to replace the existing use of the DOM, 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* [8], *XDuce* [9], *CDuce* [4], and *XQuery* [5]. Xquery was ultimately chosen, for several reasons: having a W3C recommendation, the available resources about the language are more widespread; there are many efficient and free implementations; high-level operators (union, document order comparison, and node selection XPath axis were useful for sentence selection and generation) ; there are several native XML databases that include XQuery processors (BaseX[9] was used to generate and store the exercises).

3.3 Rule Engine

Since the analyzed corpus (with the STRING processing chain) used to generate the exercises is approximately 165GB in size, the Hadoop[10] Map-Reduce framework for distributed processing was used. It had already been used in the previous syntactic exercises for sentence selection, using the DOM. But this required a new verbose Java program for each exercise, increasing complexity.

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

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. 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

[9] `http://basex.org` (last visited in October 2012)
[10] `http://hadoop.apache.org` (last visited in October 2012)

can be generated on-the-fly by the interface (for example, if the variation is in pronoun or 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 of 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).

4 Pronominalization Exercise

The goal of this exercise is to learn how replace a constituent by a pronoun, in a given sentence. This goal is achieved by cloze question, consisting in a stem is provided where the target constituent highlighted, and a set of alternative answers, a correct form and three incorrect forms, or distractors.

Pronouns can have tonic or atonic forms. Atonic forms are prone to cliticization, when they are moved next to a verb. 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 3^{rd} 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, 3^{rd} person pronouns *o, a, os, as* assume the form *lo, la, los, las*. In that case, the verb looses 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 on and the pronouns that replaces it takes the correspondent case.

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 morphemes (*lavá-lo-ei* "I will wash it"; *lavá-lo-ia* "I would wash it").

Example

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 **it** to Ana.) [The pronoun should be in the accusative case because the constituent is the direct complement. The correct position for the clitic is after the verb because this is a declarative, affirmative sentence, the verb the pronoun depends on is not in a sub clause and no special quantifiers on the subject nor any adverbs interfere with the clitic position.]

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 pronoun form.]
- *O Pedro **o** deu à Ana.* (Pedro **it** gave to Ana.) [Wrong clitic position.]

4.1 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 (see section 4.5).

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. These 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.

4.2 Sentence Selection

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, such as maximum word number and presence of clitics of the same case being taught. There are also filters to prevent sentences with NLP analysis errors to be proposed for generation. One example are sentences with the ambiguous word *que* (that/which), which in many cases introduces a subclause. However, parsing errors sometimes ignore the subclause status, introducing errors in the exercise generation. Such sentences are filtered. Other filters apply to each phase of the generation, described on the following sections.

4.3 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.

Some filters were applied to the complement selection: complements have to be noun phrases; complements is subclauses should not be pronominalized; indefinite complements cannot be pronominalized; complements cannot have appositions; the complement cannot be followed by a relative clause introduced by (a facultative preposition and) *que/o qual/cujo*.

The complement dependencies in STRING only detect the head of the constituent. To recover the entire constituent, several steps were taken. The basic selection consists of including the whole node in which the complement head appears. Then, for each complement head, modifiers are added in a recursive fashion. The modifiers can be adjectives or prepositional phrases which start with *de* (of). When there is a conjunction of several complement dependencies on the same verb, they are joined. 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* (the next ministers of Defense and of Foreign Affairs), 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.

Finally, 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 GF also seized the Bedford car), 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.* (It is important, above all, the concept of serving the user in a effective way), 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.*(The exports would serve to justify the exit of the materials bought by Joaquim Oliveira), in which the last PP was not attached to the complement as it should.

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* (half of the total investment), pronoun:*o*). Otherwise, the gender/number comes from the quantifier (ex: *fardos de* ***palha*** (straw bales), 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.* (João took Teresa and Carlos to movies) becomes *O João levou-****os*** *ao cinema.* (João took them to the movies).

4.4 Pronoun Case and Form Generation

The case is an argument of the generation and depends on the complement dependency. In the dative case, since only 3^{rd} person pronouns were considered for this exercise, so that 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.

4.5 Pronoun Positioning Rules

There are 6 rules for complement pronouning, common to both accusative and datives pronouns. All rules record generation information (e.g. for feedback purposes), such as the verb and its complement, pronoun case and position, etc.

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*. For example:

Mário Soares, por seu lado, elogiou ***a personalidade do visitante****..*
*Mário Soares, por seu lado, elogiou-****a****.*
Mário Soares, in his turn, praised ***the visitor's personality/-it****.*

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. In this exercise only verbal chains with one auxiliary verb are considered. 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.

There are 12 possible combinations of sentence types (main, negative or subclause) and clitic positions. There can be more than one correct position for each verb and feature set.

The constraints on clitic position were obtained mostly by introspection, using example sentences to derive the correct positioning for each feature set. 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.

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.*

Rule 4: Indefinite and negative subjects.

This rule deals with pronouns and determiners that modify the subject. Indefinite pronouns, e.g. *alguém* 'somebody' and negative indefinite pronouns e.g. *ninguém* 'nobody', attract the clitic pronoun to the pre-verb position. This also happens when the subject is a common noun with some quantifier determiners and some indefinite determiners. However, some of this 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.* (All boys play football) [quantifier determiner *todos* modifies the subject `NP` *os rapazes*].
- *Todos jogam à bola.* (All play football) [the subject is the quantifier determiner alone].

The `DETD` (definite determiner) and `QUANTD` (quantifier determiner) syntactic 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.

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.

Rule 6: Subordinate clauses.

In subordinate clauses, clitics are attracted to pre-verbal position. This takes place in completives, relatives and adverbial subordinate clauses; it is also a feature of direct partial interrogatives.

4.6 Distractor Generation

There are four types of distractors: wrong case, wrong position, combination of wrong case and position, and wrong accusative form distractors.

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, and the generation needs syntactic information available in that phase. However, the accusative case form distractors are generated during the presentation, by the removal or addition of one character in the clitic from the correct answer.

4.7 Exercise 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 distractors. 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. A feedback interface based on templates presents explanations about the answer to the student, along with examples from the sentence, so he/she can understand and learn all the aspects pertaining to the pronominalization (case, position and form). Several grammatical explanations are also included in tool-tips that appear when the user hovers the mouse cursor over the underlined words.

5 Evaluation

5.1 Evaluation Setup

The exercises were generated from the CETEMPublico [15] newspaper corpus, that includes approximately 8 million sentences, according to its official website [11]. 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. For all sentences, 1,292,888 exercises were generated, and 206,967 exercises for sentences with less than 20 words.

The evaluation of exercises generated from the corpus cannot encompass all generated exercises, as the number of generated exercises is too large for manual inspection. 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. A total of 240 exercises (20 for each of the 6 rules) were evaluated.

Precision was chosen as the evaluation measure, defined as the number of correct exercises by the total number of evaluated exercises.

A website was made available for testing by both native speakers and non-native Portuguese students (Fig. 3).

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 4.5). One of the factors to be analyzed was the nature of

[11] `http://www.linguateca.pt/CETEMPublico` (last visited in October 2012).

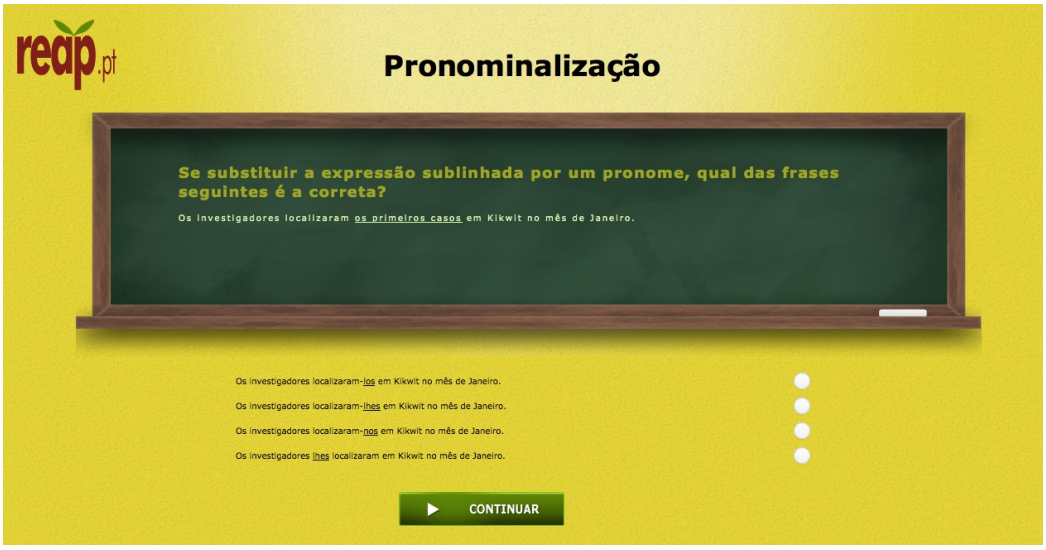


Figure 3 REAP.PT new syntactic exercises interface.

the errors that are committed by speakers of different levels, namely the distractor type in the wrong answers.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.

5.2 Evaluation Results

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.

For each incorrect exercise, the error causes were annotated by the expert. The following causes were found: PP-attachment problem (in the complement delimitation); *verbum dicendi* (incorrect identification of the inverted subject in a *verbum dicendi* construction); wrong clitic positioning; incorrect POS tagging; incorrect attachment of the pronoun to the verb; and other (corpus errors, fixed expressions, etc.).

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 mesoclisis). The PP-attachment problem was the most prevalent, with 44% of the incorrect exercises. The linguistic information in the corpus analysis is not sufficient to solve this problem.

Crowd-sourced Test Results.

The native speakers (NS) results were obtained from 114 users, with an average age of 31.5, ranging from 18 to 61 years old. The non-native speakers (NNS) results were obtained from 19 users, with an average age of 31.8, ranging from 20 to 60 years old.

For NS, main clauses had the fewest incorrect answers (10.9%), being the simpler sentences. While verbal chains have the most complex structures and rules, they do not exhibit a higher error percentage than average (20.8%). The highest number of incorrect answers, for both NS (50.5%) and NNS (33.3%) happens with sentences that have indefinite subject (pronouns or determiners). These sentences also happen to be the ones with more exercises deemed erroneous by the users. For NNS, the incorrect answers appear uniformly distributed among the positioning rules, with an average of 29%. Clauses with adverbs had the fewest incorrect answers.

The distribution of incorrect answers by distractor type was also analyzed. For NS, most errors occur with position distractors (45.5%), 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 (27.9%). For NNS, the position and case combination 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%).

Questionnaire Results.

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

The statement about exercise difficulty had less agreement between evaluation subjects. 38% of the NS and 13% of the NNS thought the difficulty was acceptable; 37% of NS and 40% of NNS disagreed, noting that the exercises may be difficult. On the other side, 26% of NS and 47% of NNS agreed that the exercises were too easy.

The majority of the users also agreed that the feedback (Fig. 4) was sufficient explanation for the answers. None of the NNS disagreed, compared to the 6% NS that found the feedback could be more detailed, or with more examples as seen in the comments.

More notably, 71% of the NS and 80% of the NNS agreed or strongly agreed that the system is useful and they learned something by using it. 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 vast majority (85% for both groups) were somewhat or very satisfied.

Questionnaire Comments.

In the free-form text comments at the end of the questionnaire, several problems were raised and suggestions were made. The most common were about the lack of context for some sentences, and the complexity of the feedback explanations (on the other hand, many praised the feedback system).

6 Conclusion and Future Work

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.

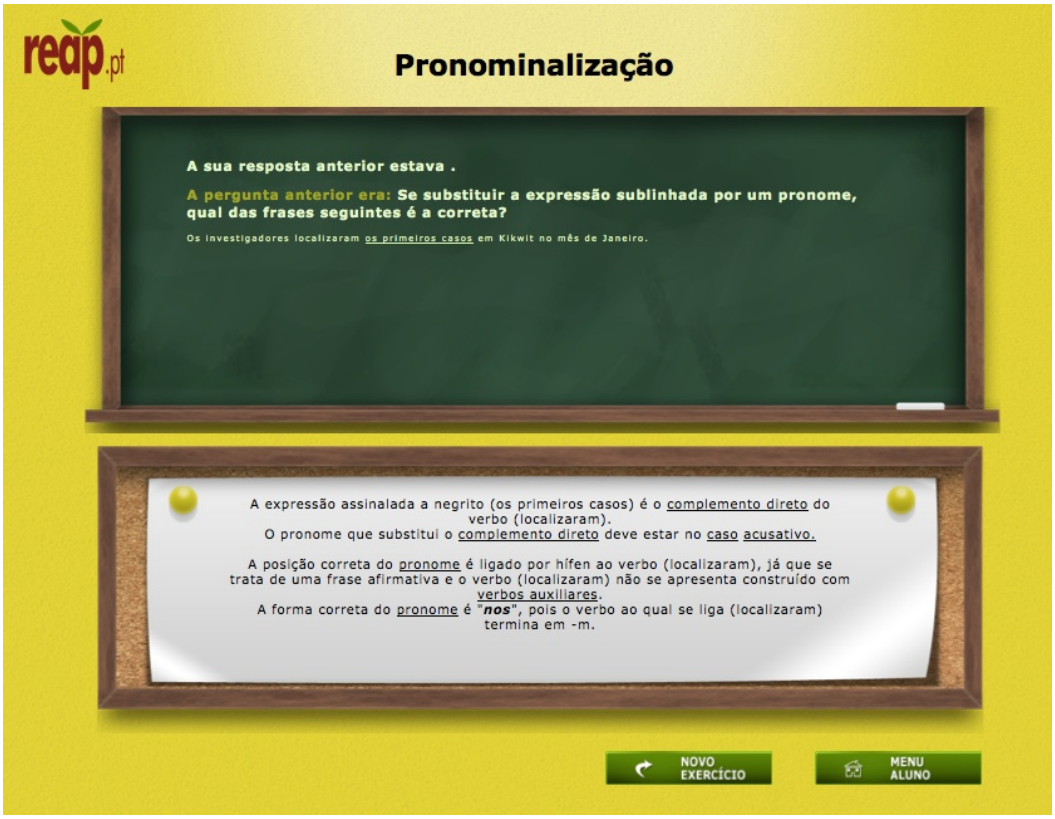


Figure 4 REAP.PT new syntactic exercises interface.

We believe that this work is a valuable new asset for the creation of new syntactic exercises for the European Portuguese language. The general architecture of the REAP.PT syntactic module is expected to make a relevant step forward in order to ease the development effort of 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.

Some pitfalls were also uncovered during the development, such as the unapparent complexity of some aspects of syntactic exercise generation, and the heavy reliance on correctness and completeness of the NLP analysis of the text. Therefore, the analysis of the exercise generation approach and NLP analysis of the information needs are very important for the success of this exercise's 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, 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.

Regarding the future work, the errors detected during the evaluation should be corrected; the future-indicative and conditional tenses should be implemented; and exercises could be generated from other corpora, to add variety.

References

1 S. Aït-Mokhtar, J.-P. Chanod, and C. Roux. Robustness beyond shallowness: incremental deep parsing. *Nat. Lang. Eng.*, 8(3):121–144, June 2002.

2 Itziar Aldabe. *Automatic Exercise Generation Based on Corpora and Natural Language Processing Techniques*. PhD thesis, Euskal Herriko Unibertsitatea (University of the Basque Country), San Sebastian, Basque Country, September 2011.

3 Itziar Aldabe, Maddalen Lopez de Lacalle, Montse Maritxalar, and Edurne Martinez. The Question Model inside ArikIturri. In J. Michael Spector, Demetrios G. Sampson, Toshio Okamoto, Kinshuk, Stefano A. Cerri, Maomi Ueno, and Akihiro Kashihara, editors, *Proceedings of the 7th IEEE International Conference on Advanced Learning Technologies, ICALT 2007, July 18-20 2007, Niigata, Japan*, pages 758–759. IEEE Computer Society, 2007.

4 Véronique Benzaken, Giuseppe Castagna, and Alain Frisch. CDuce: an XML-centric general-purpose language. *SIGPLAN Not.*, 38(9):51–63, August 2003.

5 Don Chamberlin. XQuery: a query language for XML. In *Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data*, SIGMOD '03, pages 682–682, New York, NY, USA, 2003. ACM.

6 Chia-Yin Chen, Hsien-Chin Liou, and Jason S. Chang. FAST: an automatic generation system for grammar tests. In *Proceedings of the COLING/ACL on Interactive Presentation Sessions*, COLING-ACL '06, pages 1–4, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

7 Rui Correia. Automatic Question Generation for REAP.PT Tutoring System. Master's thesis, Instituto Superior Técnico - Universidade Técnica de Lisboa, Portugal, 2010.

8 B. Emir. Extending pattern matching with regular tree expressions for XML processing in Scala. Master's thesis, RWTH Aachen, 2003.

9 Haruo Hosoya and Benjamin C. Pierce. XDuce: A statically typed XML processing language. *ACM Trans. Internet Technol.*, 3(2):117–148, May 2003.

10 Nuno Mamede, Jorge Baptista, and Caroline Hagège. Nomenclature of Chunks and Dependencies in Portuguese XIP Grammar 3.1. Technical report, L2F/INESC-ID, Lisbon, May 2011.

11 Nuno J. Mamede, Jorge Baptista, Cláudio Diniz, and Vera Cabarrão. STRING: An Hybrid Statistical and Rule-Based Natural Language Processing Chain for Portuguese. `http://www.propor2012.org/demos/DemoSTRING.pdf`, April 2012.

12 Cristiano Marques. Syntactic REAP.PT. Master's thesis, Instituto Superior Técnico - Universidade Técnica de Lisboa, Portugal, 2011.

13 Luís Marujo. REAP em Português. Master's thesis, Instituto Superior Técnico - Universidade Técnica de Lisboa, Portugal, 2009.

14 Thomas Pellegrini, Rui Correia, Isabel Trancoso, Jorge Baptista, and Nuno J. Mamede. 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*, pages 1629–1632. ISCA, 2011.

15 Diana Santos and Paulo Rocha. Evaluating CETEMPublico, a Free Resource for Portuguese. In *Association for Computational Linguistic, 39th Annual Meeting and 10th Conference of the European Chapter, Proceedings of the Conference, July 9-11, 2001, Toulouse, France*, pages 442–449. Morgan Kaufmann Publishers, 2001.

16 Alberto Simões and Diana Santos. Ensinador: corpus-based portuguese grammar exercises. *Procesamiento del Lenguaje Natural*, 47:301–309, September 2011.