# Automatic Generation of Exercises on Passive Transformation in Portuguese

Jorge Baptista
INESC-ID Lisboa
Faculdade Ciências Humanas e Sociais
Universidade do Algarve, Portugal
jbaptis@ualg.pt

Sandra Lourenço
INESC-ID Lisboa
Instituto Superior Técnico
Universidade de Lisboa, Portugal
Sandra.Lourenco@tecnico.ulisboa.pt

Nuno J. Mamede
INESC-ID Lisboa
Instituto Superior Técnico
Universidade de Lisboa, Portugal
Nuno.Mamede@tecnico.ulisboa.pt

*Abstract*—**Technology plays a very important role in education and Intelligent Computer-Assisted Language Learning (iCALL) has emerged as a complementary or even alternative method to the conventional language teaching practices. The automatic generation (and correction) of language exercises based on real texts extracted from corpora constitutes a non-trivial challenge to iCALL tutorial systems, and may involve the use of sophisticated Natural Language Processing tools and large-scale linguistic resources. This paper presents the main issues related to the automatic generation of exercises on the Passive transformation, a commonly occurring type of exercises in language textbooks, but also a very complex topic of Portuguese grammar. The paper describes the methods used to produce a large batch of passive-active sentence pairs, where the active sentence was automatically generated from naturally occurring passive sentences, taken from a large-sized, publicly available, corpus. Sentence pairs are ranked by difficulty level. A sample of randomly selected sentence pairs (40 from difficult level, 100 from medium, and 100 from easy level) was manually evaluated by an expert. Results are presented and error analysis is performed. The sentence pairs can be used as prime and correct answer for iCALL systems.**

## I. Introduction

With the increasing use of technology, the traditional language teaching and learning paradigms are evolving to a more remote-access, globalized, and computer-assisted environment, involving new practices and methods. Intelligent Computer Assisted Language Learning (iCALL) has thus emerged as a complementary, or even an alternative, method to conventional student-teacher interaction in language teaching.

The main goal of this work is to extend the functionalities already available in the REAP.PT [9], [11], [14], [16], [19], an iCALL tutorial system for Portuguese, by developing a new module of exercises focused on the transformation of sentences from the passive to the active construction.

This grammatical topic is frequently present in traditional textbook exercises and is taught in Portuguese classes, starting in the sixth grade and being continuously taught until the twelfth grade. The passive transformation in Portuguese can involve really complex lexical-syntactic constraints and it entails several morphosyntactic changes from one sentence form to the other. Besides, the automatic generation of active sentences from their passive counterparts poses non trivial challenges to NLP systems. Given its importance in the Portuguese curriculum, it is important that there be tools and resources that automatically generate and evaluate this kind of exercises, so that students can practice them autonomously, thus saving an important work load to language teachers.

This paper is structured as follows: Section II presents related work, particularly existent CALL and iCALL systems supporting Portuguese teaching and learning. In Section III a brief overview of the Passive transformation in Portuguese is presented. Next, Section IV describes the sentence selection process and the morphological and syntactical changes performed by the algorithm that generates the active sentences. The evaluation and detailed error analysis are presented in Section V. Finally, Section VI concludes the paper and draws a map for future work.

## II. Related work

In this section, we present previous work related to the topic of the paper. First, we provide a quick overview of existing websites with interactive exercises for Portuguese learning practice, followed by several currently available CALL systems, in order to better frame the context of the system where the exercises on Passive are to be implemented.

Table I presents an overview of the sites and CALL systems currently available for learning Portuguese, their public availability and whether the exercises and their feedback are automatically generated or not.

### *Language teaching resources*

*Ciberescola da Língua Portuguesa*[1] is a freely accessible, online platform for Portuguese teaching. It proposes exercises for Portuguese native-speakers from the 5th to the 12th grade, and also exercises for students of Portuguese as a Second Language, for levels A2 to B1, classified according to the *Common European Framework of Reference for Languages* [10]. These exercises cover reading and oral comprehension, grammar, writing skills and vocabulary training, and are organized in three levels of difficulty: easy, normal and hard. Exercises are manually produced by teachers and linguists, and they are also scored. The number os exercises is static, and students can track the number of exercises left to be solved. Feedback consists of providing the user with the correct answer and

[1] www.ciberescola.com (last access: January 2016).

978-1-5090-0623-6/16/$31.00 ©2016 IEEE

TABLE I
CALL SYSTEMS OVERVIEW

| | Publicly Available | Automatically Generated | Dynamic Feedback |
|---|---|---|---|
| Ciberescola da Língua Portuguesa | X | | |
| Centro Virtual Camões | X | | |
| Observatório da Língua Portuguesa | X | | |
| TAGARELA | | | X |
| WERTi | X | X | X |
| The Alpheios Project | X | X | X |
| FAST | | X | X |
| ArtikIturri | | X | X |

her/his score. Ciberescola has three types of exercises about active and passive sentences, all involving auxiliary verb *ser* 'be'. The first type (targeted at the $6^{th}$ grade) consists only in identifying whether a sentence is in the active or passive form. The second ($7^{th}$ grade) involves producing the passive sentence from an active prompt, but by merely filling blank spaces. In the third type of exercises ($8^{th}$ grade), the student has to relate/link the active with the correspondent passive.

*Centro Virtual Camões*[2] is part of Instituto Camões, a Portuguese public institute with the mission of promoting the Portuguese language and culture worldwide. This web page provides language-related games as well as reading and listening exercises, aimed at several aspects of history, culture, grammar and vocabulary. All the games and the exercises have a difficulty level assigned in a three-level scale. Centro Virtual proposes three activities that involve exercises about the Passive transformation. In all of them, the student has to transform active sentences into their passive counterpart, by filling in blank spaces. Only passive sentences with auxiliary verb *ser* 'be' are considered.

*Observatório da Língua Portuguesa*[3] is a non-profit organisation whose objective is the promotion of the Portuguese language. The webpage has some Portuguese learning resources available, namely some videos and multiple-choice or fill-in-the-blanks exercises, both aiming at lexical and grammatical competences. Those exercises are not automatically generated and there is no available correction nor feedback after the answers are submitted. To date, no exercises about active-passive are provided.

### *CALL systems overview*

*TAGARELA* (Teaching Aid for Grammatical Awareness, Recognition and Enhancement of Linguistic Abilities)[4] [2] is an iCALL system for Portuguese, whose development started at the Ohio State University and was later moved to the University of Tübingen. TAGARELA is divided in two components, which perform the syntactic and semantic analysis of the users' answers. The first component, Form Analysis, consists of four modules: tokenizer (with special attention to cliticization, contractions and abbreviations); lexical/morphological lookup (returns multiple analysis based on the CURUPIRA lexicon [15]); (rule-based) disambiguator; parser ( bottom-up syntactic analyzer). The second component, Content Analysis, tries to establish the semantic agreement between the user's answer and the correct answer. This system can be accessed through a web browser and uses NLP technology to analyze the users' input and detect its spelling, morphological, syntactic and semantic errors. The exercises available are similar to the ones found in textbooks and include: listening and reading comprehension, picture description, vocabulary practice, phrasing and re-writing. TAGARELA keeps track of the users' performance based on the previously solved exercises and provides personalized feedback based on the users' errors in each exercise. The system, however, does not provide automatically generated grammatical (syntactic) exercises.

*WERTi* (Working With English Real-Texts)[5] [17] is an iCALL system designed to help students that are learning English as a second language (Spanish and German working modules are in a beta phase). Like TAGARELA, its development started at Ohio State University and was later moved to University of Tübingen.

The exercises present in WERTi are generated from real texts, selected from the web by the user using a a toolbar plugin that can be installed in Firefox. An NLP chain processes these texts on-the-fly at the server's side and identifies specific targets within those texts to generate the exercises.

The processing is done in three phases: HTML processing (pre-processing: removes the content of the page that surrounds the text, returning only the relevant text); linguistic processing (responsible for tokenization, sentence boundary detection and POS tagging); and enhancement of processing (post-processing; performs the necessary operations for the creation of the exercise).

This system provides exercises about several grammatical topics: articles, other determiners, gerunds vs. infinitives, phrasal verbs, prepositions and Wh-questions. Regardless of the topic chosen by the user, she/he can choose one of the three following activities:

- Colorize - The linguistic forms that the student is supposed to learn are highlighted;
- Click - The user has to click on the answers she/he thinks are correct;

[2] cvc.instituto-camoes.pt (last access: January 2016).

[3] observatorio-lp.sapo.pt/pt (last access: January 2016).

[4] sifnos.sfs.uni-tuebingen.de/tagarela (last access: January 2016).

[5] sifnos.sfs.uni-tuebingen.de/WERTi (last access: January 2016).

- Practice - The user is asked to perform different kinds of exercises, such as filling in the blanks.

Almost any combination of type and format of exercise is available. In all of the cases, feedback is provided by coloring the user's answers: correct answers are colored green and incorrect answers are colored red. WERTi does not have any exercises about active and passive sentences.

The Alpheios Project[6] is an open-source project developed by The Alpheios Project non-profit organization. The main focus of this project is on ancient languages, such as Latin and ancient Greek. Arabic is also available and support for classical Chinese is under development. Plans for the future include the support of Sanskrit, Hebrew and Persian. To support the learning of the already available languages, the system provides features such as word definitions, word morphology, translation, inflection tables, sentence diagrams, quizzes and a personal word list manager. Those features can be used either on texts uploaded by the user or on texts already preprocessed by the system. The Alpheios Project does not have any exercises about active and passive sentences.

FAST (Free Assessment of Structural Tests) [7] is a system that automatically creates English grammar tests. The test items are created by gathering sentences from the Web, which are based on test patterns adopted from the Test Of English as Foreign Language (TOEFL). With this method, test writing knowledge is represented as structural patterns, acquiring real sentences on the Web and applying patterns to transform those sentences into questions. The questions can be of two types: multiple-choice and error detection.

ArikIturri [3] [4] is a system that automatically generates different types of questions, such as: fill-in-the-blank, word formation, multiple-choice, error correction and short answer. Those questions can be generated in two languages: Basque and English. It was developed at the IXA research group[7], at the University of Basque Country. The exercises in ArikIturri are a set of questions composed of the following elements: topic, answer focus and context. The answer focus 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 (which contains the necessary information to treat the topic) and a notHead. The head is divided into three elements: answer, 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. The distractors are linked to the answer focus because they are the list of words that are incorrect in that context. The headComponent collects the information related to the question type [4]. This question model is general and flexible. It is considered general because it is independent from the language in which the questions are generated and also from the NLP tools used for the generation. It is considered flexible because it foresees that new types of questions can be added and permits to have more than one answer focus into the same question [4].

### *The NLP system STRING*

For collecting candidate sentences that will be the primes of the exercises on passive sentences, the freely available CETEMPúblico corpus [8] was used. This corpus is composed of extracts taken from the online edition of the daily newspaper *Público*, totaling over 190 million words. It had been previously processed using the STRING processing chain [13][8] and its xml output was used to extract the candidate sentences. This section briefly presents the system's main features used for this task.

STRING is a hybrid, statistical and rule-based, natural language processing chain developed in a modular architecture for processing Portuguese texts. It performs all the basic NLP task: text segmentation, POS tagging, statistical and rule-based POS disambiguation, shallow parsing (chunking) and syntactic analysis (deep parsing). For the chunking and parsing steps, STRING uses the Xerox Incremental Parser (XIP) [1], a rule-based finite-state parser that allows for the identification of elementary constituents (or *chunks*, *e.g.* NP, PP, etc.) and the extraction of the syntactic dependencies between the chunks' heads (*e.g.* SUBJ[ect], CDIR (direct complement), MOD[ifier], etc.).

Certain auxiliary dependencies, like DETD (determinant dependency) or POSS (possessive determiner) for example, link elements within a given constituent, *i.e.* the determiner and the noun it determines. Dependencies can take features such as _PRE or _POST, depending on the relative position of the governor and governed element. By default, PP are first linked to another element by the umbrella MOD dependency, but if their status as essential argument of a verb is determined, this dependency is replaced by COMPL[ement]. This is the case of the agentive PP complement of passive sentences. Portuguese has a very complex system of auxiliary verbs, conveying aspect, modality and even tense values that the verb inflection alone can not express. Auxiliaries can recursively combine themselves to produce complex verbal chains. The relation between an auxiliary verb and the verb it operates on is captured by auxiliary dependency VLINK. A VDOMAIN dependency delimits the first and last (or main) verb of a verbal chain.

STRING takes a text as input and outputs an xml file, where each node has a set of attribute-value pairs with the linguistic information produced. It is from this xml file that sentences have been extracted to produce the passive-active exercises.

## III. Types of Passive Sentences

In the case of *action* direct-transitive verbal predicates, the subject of an active sentence has the semantic role of *agent*, while the direct complement is the *object* or the *patient* of the action, that is, the element that undergoes the action (examples taken from [18]): *Pedro Nunes inventou o nónio.* (Pedro Nunes invented the vernier). Passive sentences are characterized mainly by having a different alignment than that from active sentences: *O nónio foi inventado por Pedro Nunes.*

[6] alpheios.net (last access: January 2016).

[7] ixa.si.ehu.es/Ixa (last access: January 2016).

[8] string.l2f.inesc-id.pt (last access: January 2016).

(The vernier was invented by Pedro Nunes). Both sentences have the same sentence type (declarative, affirmative), describe the same situation, keep unchanged the same semantic roles for the verb's arguments, and are characterized by the same illocutory force.

Five types of passive sentences can be considered:
**(i)** *Verbal passive sentences* are sentences in which occurs a complex verbal group started by the auxiliary verb *ser* (be) followed by a participle corresponding to the verb of the active sentence; they typically describe situations in which one of the entities involved goes through some kind of change (example above);
**(ii)** *Resultative passive sentences* describe a situation that is the result of a change in state, place or possession; in this kind of passive sentences, it is the verb *ficar* (become/stay) that usually occurs: *Os museus de Bagdade ficaram destruídos (como resultado dos bombardeamentos americanos).* (The Baghdad museums were destroyed (as a result of the American bombing)) (← *Algo destruiu os museus de Bagdade*) (Something destroyed the Badgad museums);
**(iii)** *Stative passive sentences* are passive sentences that are syntactically close with the passive resultative sentences but instead of focusing on the result of the change, they focus on describing stative situations whose meaning does not have anything related with the change of state: *Este autor é muito conhecido.* (This author is well known) (← *Alguém/As pessoas conhece(m) este autor*) (Someone/People know this author);
**(iv)** In *pronominal passive sentences* the passive is expressed through reflexive pronouns with the object aligned in subject position: *As camisas lavam-se sempre a frio.* (The shirts are always washed at cold temperature) (← *Alguém/As pessoas lavam sempre as camisas a frio*) (Someone/People always wash the shirts at cold temperature);
**(v)** *Neutral* or *median passive sentences* are sentences where the *object* (or *theme*) appears aligned in the subject position; the active sentence subject has a semantic role of *cause* and is either dropped or is introduced by proposition *com* (with) or other causal connectors, (e.g. *por causa de* (because of)): *As condutas entupiram (com o lixo).* (The pipes clogged (with/because of the garbage))(← *O lixo entupiu as condutas*) (The garbage clogged the pipes).

In this paper, we focus on verbal passives alone, with auxiliary verb *ser* (be). Furthermore, we present a method to generate active sentences from their passive counterparts and not the other way around, as it is usual the case in much of the available on-line and textbook exercises.

## IV. Implementing Passive to Active Transformation

To automatically generate exercises that perform the transformation sentences from the passive to the active structure, some preliminary steps are necessary:

A. selection of the passive sentences that will be used as *primes*;
B. morphological transformation of the passive verb; and
C. syntactical transformation of the sentence itself.

Accessorily, exercises will be rank by difficulty level, being necessary to define criteria to automatically rank them. As mentioned above, the sentences used in this work come from the CETEMPúblico corpus [8], which had been previously processed using the STRING processing chain [13].

### A. Sentence Selection

The following constraints are used to ensure that the selected sentences have the best possible quality to be used for educational proposes:

Since one aims at generating an active from the corresponding passive sentence, both the *agent* and the *object* must be present in the passive sentence that is to be extracted and serve as prime; only one dependency of type `COMPL_POST` (representing the passive *agent* and only one dependency of type `SUBJ_PRE` (representing the passive sentence subject) can be present (for the same main verb), in order to ensure that the sentence only has one passive verbal chain; only one dependency of type `VDOMAIN` (representing the passive verb) is allowed; no dependencies of type `POSS` (that represent possessive determiners) are allowed; the agentive complement is the prepositional phrase (`PP`) that starts with the preposition *por* (by) and its head noun must be a human noun; the number of verbs connected by the `VLINK` dependency is different for each difficulty level; the main verb in the verbal chain captured by the `VDOMAIN` dependency has to be part of the list of verbs that accept that passive transformation with the auxiliary verb *ser* (be); this information is retrieved from the verbs' lexicon, *ViPEr* [6] (currently, 4,254 out of 6,640 (syntactic-semantic) verb entries (5,127 different lemmas) allow the verbal passive with this auxiliary).

Besides the length of the verbal chain, the sentence difficulty level is also defined by the number of dependency relations extracted by the parser. A sentence is considered to belong to the *easy* level if it has only one `VLINK` dependency and less than 55 dependency tags in the `xml` corpus file. There were 6,549 sentences found. The *medium* level includes 1,364 sentences with two `VLINK` dependencies (in the same `VDOMAIN`) and less than 65 dependency tags in the `xml` corpus' file. Finally, sentences considered to belong to the *difficult* level feature three `VLINK` dependencies and less than 75 dependency tags in the `xml` corpus' file: 80 sentences found.

### B. Morphological Verb Transformation

In the easy level, the verbal chain has only two verbs: the auxiliary verb (the passive auxiliary *ser* (be)); and the main verb, in the past participle form; there is only one `VLINK` dependency. It is the main verb that will be transformed to the active form, taking into account the tense of the verb *ser* (be). The transformation is done with the help of LexMan [20], by searching for the form of the verb in the past participle that correspond to the same tense and mood as the tense and mood of the auxiliary verb *ser* (be), and in the same number as the noun that was identified as the sentence's *agent*. Example: *O livro* ***foi***$_{PastPerf}$ ***lido***$_{PastPart}$ *pelos alunos*$_{pl}$ (The book was read by the students) → *Os alunos*$_{pl}$ ***leram***$_{PastPerf-pl}$ *o livro.* (The students read the book).

In the medium difficulty level, the verbal chain has three verbs. The second verb is always the passive auxiliary *ser* (be) and the third verb is always the main verb (in the past participle form); there are two `VLINK` dependencies; the first verb of the chain will stay in the same tense, but may need to be changed (from the singular to the plural form, or the opposite) in order to agree with the agent noun. The main verb will be transformed to the active form, with the help of LexMan, by searching for this verb in the tense of the second auxiliary verb in the sentence. The main verb will be in the same tense as the verb *ser* (be) in the passive sentence. Example: *O livro* ***tinha***$_{PastImperf}$ ***sido***$_{PastPart}$ ***lido***$_{PastPart}$ *pelos alunos*$_{pl}$ (The book had been read by the students) → *Os alunos*$_{pl}$ ***tinham***$_{PastImperf-pl}$ ***lido***$_{PastPart}$ *o livro* (The students had read the book).

In the difficult level, the verbal chain has four verbs; the third verb is always the passive auxiliary *ser* (be); there are three `VLINK` dependencies; the first verb will be transformed in the same way as in the medium level; the second verb is usually in the infinitive form and will stay the same in the active sentence; the main verb will be transformed to the active form in the same way as it is in the medium level. Example: *O livro* ***podia***$_{PastImperf}$ ***ter***$_{Inf}$ ***sido***$_{PastPart}$ ***lido***$_{PastPart}$ *pelos alunos*$_{pl}$ (The book might have been read by the students) → *Os alunos*$_{pl}$ ***podiam***$_{PastImperf-pl}$***ter***$_{Inf}$ ***lido***$_{PastPart}$ *o livro* (The students might have read the book).

### C. Syntactical Transformation from Passive to Active

Not all sentences' transformations are straightforward and some additional steps may have to be performed. In some passive sentences, words appearing inside the verbal chain may have to be moved to an equivalent position in the active sentence. Example: (a) *O discurso parece ter sido* ***bem*** *entendido pelos destinatários.* (The speech seems to have been **well** understood by the recipients.) → (b) *Os destinatários parecem ter entendido* ***bem*** *o discurso.* (The recipients seem to have understood **well** the speech.). In this case the adverb *bem* (well) is moved to after the main verb (cp. */?* *Os destinatários parecem ter* ***bem*** *entendido o discurso.*).

If the passive sentence starts with a connective adverb, usually that word has to stay in the beginning of the active sentence. The rest of the sentence will be transformed following all the basic rules. Example: (a) ***Mas*** *a proposta do coordenador da Comissão Permanente do PS foi recusada pelos sociais-democratas.* (**But** the proposal of the coordinator of the Permanent Commission of the PS (=Socialist Party) was rejected by the social-democrats) → (b) ***Mas*** *os sociais-democratas recusaram a proposta do coordenador da Comissão Permanente do PS.* (**But** the social democrats rejected the proposal of the coordinator of the Permanent Commission of the PS) Notice that, while *mas* (but) is typically a coordinative conjunction, in this case it is used in much the same way as a connective adverb, like *porém* (however).

When transforming passive sentences that are in the negative form, special care is needed to ensure that the active sentence also remains in the negative form and that the words that identify the negative sentence are put in their correct places. Example: (a) *A administração da RTP* ***não*** *pode continuar a ser nomeada pelos governos.* (The board of direction of RTP (=Public Television Network) **cannot** continue to be nominated by governments) → (b) *Os governos* ***não*** *podem continuar a nomear a administração da RTP.* (Governments **cannot** continue to nominate the board of direction of RTP). In the case of indefinite-negative determiner *nenhum* (no/none), when the negation is part of the subject of the passive sentence, the active counterpart involves the insertion of the negation adverb *não* (no): (a) *Porém,* ***nenhuma*** *decisão concreta foi tomada pelos britânicos em Toulouse.* (However, **no** concrete decision was taken by the British in Toulouse) → (b) *Porém, os britânicos em Toulouse* ***não*** *tomaram* ***nenhuma*** *decisão concreta.* (However, the British in Toulouse took **no** concrete decision).

The next aspects concern formal aspects and punctuation, which can hinder results if inadequately handled.

Some passive sentences have a comma or colon somewhere in the first part of the sentence (before the verbs). In most of the cases, the words behind the comma (and the comma itself) have to remain in the same place in the final sentence (the same for sentences with colon). The rest of the sentence will be transformed as usual. Example: (a) ***Na recta final,*** *o empresário foi ultrapassado, em capacidade financeira, por António Champalimaud.* (**In the final stretch,** the businessman was surpassed in financial capacity, by António Champalimaud) → (b) ***Na recta final,*** *António Champalimaud ultrapassou o empresário, em capacidade financeira.* (**In the final stretch,** António Champalimaud surpassed the businessman in financial capacity).

In some cases, the passive sentences have fractional numbers. In Portuguese orthographic convention, the decimal separator is the comma <`,`>, while the thousands' separator is the dot <`.`>. This has to be taken into account to ensure that the numbers from both sides of the decimal separator stay together in the correct side of the sentence.

When the passive sentence is enclosed by parenthesis, it is necessary to make sure that those parenthesis do not end up in the middle of the active sentence and that the active sentence is also enclosed by the parenthesis. The same thing has to be done when the whole passive sentence is enclosed by quotes. Example: (a) ***"****O sistema de alvarás tem sido regulado pelos empreiteiros****"****.* (**"**The permits system has been regulated by contractors**"**) → (b) ***"****Os empreiteiros têm regulado o sistema de alvarás****"****.* (**"**The contractors have regulated the permits system**"**). When only part of the passive sentence is enclosed by parenthesis, it is necessary to make sure that the part between the parenthesis is also between the parenthesis in the active sentence. The same thing has to be done when there is some part of the passive sentence enclosed by quotes. Example: (a) *O estudo foi desenvolvido pelo Departamento de Planeamento Estratégico* ***(DPE)*** *da autarquia.* (The study was developed by the Department of Strategic Planning **(DPE)** of the municipality) → (b) *O Departamento de Planeamento Estratégico* ***(DPE)*** *da autarquia desenvolveu o estudo.* (The Department of Strategic Planning **(DPE)** of the municipality developed the study).

At the very end of the process, the final sentence may need to be corrected: since all the contractions in the original passive sentence taken from the corpus had been previously resolved, in the final active sentence those contractions have to be reconstructed. This is facilitated by the fact that the `xml` file keeps the information of the original tokenization before the text processing.

## V. Evaluation

The evaluation was performed on a random sample of sentences taken from each of the three difficulty levels. This sample consisted of 100 sentences from the easy level, 100 sentences from the medium level and 40 sentences from the difficult level. The number of sentences from the difficult level is smaller because less sentences from this level were found in the corpus than from the other two levels. Each pair of passive-active sentences was evaluated by an expert, who determined whether the sentences that had errors and identified the nature of the errors present. For clarity, results will next be presented by level of difficulty.

### A. Results

From the *easy level*, 100 sentences were evaluated, 37 were found to have errors. Of the 63 correct sentences, 3 of those, while correct, could be transformed to the active sentence in a more natural way. This means that 60% of the sentences are transformed correctly from the passive to the active form. However, 13 errors were due to problems on the STRING procession chain and not to the passive-to-active transformation algorithm. If one discards those sentences, then 68% of the sentences had been transformed correctly. In a similar way, from the *medium level*, also 100 sentences were evaluated, 49 were errors and 8, though grammatically correct, could have been transformed in a more natural wording, hence only 43% were deemed correct. No errors due strictly to the NLP processing chain were found.

Finally, from the 40 sentences evaluated of the *difficult level*, 13 were found to have errors and 3 of those were unnatural, resulting in 60% correct results. Since 4 errors were due to the processing, this means 66% correct sentences.

Table II presents the results from this evaluation process. In this table, **S** is the number of processed sentences, **I** are the incorrectly generated sentences, **U** the unnatural and **NLP** is the number of errors due to the NLP processing chain; **P1** is the overall percentage of correct results, while **P2** is the percentage of correct results if the NLP errors were discarded.

### B. Error Analysis.

In this section we present a more detailed comment on the systems' errors. For clarity, examples for each passive-active pair are numbered and presented in the same order: "1" is the source passive sentence, "2" the incorrect sentence and, if necessary, "3" presents the expected correct result.

TABLE II
Results from manual evaluation of generated sentences

| **Level** | **S** | **I** | **U** | **P1** (%) | **NLP** | **P2** (%) |
|---|---|---|---|---|---|---|
| **Easy** | 100 | 37 | 3 | 60 | 13 | 68 |
| **Medium** | 100 | 49 | 8 | 43 | 0 | 43 |
| **Difficult** | 40 | 13 | 3 | 60 | 4 | 66 |

Coordination is a very hard parsing problem and it has significantly impacted on the results. One of the most common errors found happens when the past participle in the passive sentence appears coordinated with another one. The active sentence is incorrect as the NLP parsing system does not contemplate this type of coordination yet. Example: 1. *Este foi* ***julgado e condenado*** *pelo Tribunal de Círculo de Matosinhos a três anos de prisão.* (This_one was **tried and convicted** by the Matosinhos District Court to three years in prison). → 2.****E condenado*** *o Tribunal de Matosinhos a três anos de prisão julgou este.* (**And condemned** the Matosinhos District Court to three years in prison tried this_one.)

A similar problem occurs with complex coordinated agent complements, as only the first `PP` element keeps the proposition *por* (by) and the remaining are parsed as `NP`, thus coordinating rules are not propagated to these `NP` from the initial `NP`: Example: 1. *O mercado é feito* ***pelos produtores, distribuidores e exibidores***. (The market is made by the **producers, distributors and exhibitors**) → 2. **Os* ***produtores*** *fazem o mercado,* ***distribuidores e exibidores***. (The **producers** make the market, **distributors and exhibitors**). 3. *Os* ***produtores, distribuidores e exibidores*** *fazem o mercado.* (The **producers, distributors and exhibitors** make the market). In this example, the full agent PP includes *distribuidores e exibidores* (distributors and exhibitors).

In other cases, a sentence was identified as a passive sentence even though it is not. This can happens due to a phenomenon of length permutation leading to subject-predicate inversion. Example: 1. *Todavia, não foram muitas as saídas deixadas pelos socialistas para uma remota resposta de Cavaco.* (However, there were not many exits left by the Socialists to a remote response from Cavaco). Because of that permutation, the sequence *foram deixadas* has been parsed as passive structure, yielding the incorrect sentence: → 2. *Todavia, os socialistas não deixaram muitas as saídas para uma remota resposta de Cavaco.* The base order would be: *Todavia, as saídas deixadas pelos socialistas para uma remota resposta de Cavaco não foram muitas.* (However, the exits left by the Socialists to a remote response from Cavaco were not many).

In some cases, the use of commas in the passive sentence renders the active sentence incorrect. In the next case, the agent `PP` is separated from the main clause by comma, as if it were a sort of apposition. The resulting sentence 2, though interpretable, is incorrect, as this comma, in the active sentence, ends up between the direct and the indirect (dative) complement. This happens due to an error in part of the transformation algorithm. Example: 1. *O caso só foi contado às autoridades há três dias, por três sobreviventes da tribo.* (The case was only told to the authorities three days ago, by three survivors of the tribe). → 2. **Três sobreviventes da tribo só contaram o caso*, *às autoridades há três dias.* (Three survivors of the tribe only told the case**,** to the authorities three days ago). 3. *Três sobreviventes da tribo só contaram o caso às autoridades há três dias.* (Three survivors of the tribe only told the case to the authorities three days ago).

There were also some other problems related to the use of comma and the position of adverbials in the sentences, due to the transformation algorithm still incipient treatment of adverbial complements. In the next example, the temporal adverbial of sentence 1 should have been outside the scope of the transformation algorithm. Since this adverbial has been moved away from its initial position without separating it from the remainder of the clause by commas, a comma has been incorrectly placed between the main clause and the direct object, rendering the sentence non-grammatical. Example: 1. ***A partir do próximo ano lectivo****, o Externato Paulo VI, em Gondomar, vai deixar de ser administrado pelos padres Capuchinhos.* (**Starting next school year**, the Externato Paul VI, in Gondomar, will no longer be administered by the Capuchin priests). → 2. **Os padres Capuchinhos vão deixar de administrar* ***a partir do próximo ano lectivo,*** *o Externato Paulo VI, em Gondomar.* (The Capuchin priests will no longer administer **starting next school year,** the Externato Paul VI, in Gondomar). 3. ***A partir do próximo ano lectivo****, os padres Capuchinhos vão deixar de administrar o Externato Paulo VI, em Gondomar.* (**Starting next school year**, the Capuchin priests will no longer administer the Externato Paul VI, in Gondomar).

Since the transformation algorithm does not take into account relative clauses yet, all sentences with relative clauses have been transformed incorrectly. If the target passive is the relative clause, the transformation should only be applied within the the boundaries of that clause and whatever is outside of it should be kept unchanged. Example: 1. *um dos pontos em que o relato de Champalimaud é analisado por Freire Antunes.* (one of the points where Champalimaud's report is analyzed by Freire Antunes). → 2. **Freire Antunes que analisa um dos pontos no relato de Champalimaud.* (Freire Antunes that analyzes one of the points of Champalimaud's report). 3. *um dos pontos em que Freire Antunes analisa o relato de Champalimaud.* (one of the points where Freire Antunes analyzes Champalimaud's report).

Passive sentences that have apposition phrases or clauses may have errors when transformed to the active form because these appositions are hard to detect automatically and failing to detect them makes the algorithm leave them in place, away from their antecedent. Example: 1. *A informação da RTP vai passar a ser dirigida por* ***Manuel Rocha, até agora responsável pelo pelouro no Norte***. (The information of RTP will now be led by **Manuel Rocha, until now responsible for this area in the north**). → 2. ****Manuel Rocha*** *vai passar a dirigir a informação da RTP,* ***até agora responsável pelo pelouro no Norte***. (**Manuel Rocha** will now lead the information of RTP, **until now responsible for this area in the north**). 3. ***Manuel Rocha, até agora responsável pelo pelouro no Norte****, vai passar a dirigir a informação da RTP.* (**Manuel Rocha, until now responsible for this area in the north**, will now lead the information of RTP).

When there are personal pronouns in a sentence, the verb has to be transformed in a different way, taking into account those pronouns [12]. Since pronouns were not taken into account in the algorithm here developed, some sentences with personal pronouns are transformed incorrectly. Notice that this pronoun position and form are acceptable in Brazilian Portuguese. Example: 1. *Mais,* ***ela tem sido considerada****, por numerosos historiadores, eventualmente com razão, como o verdadeiro motor de civilização.* (Furthermore, **she has been considered**, by many historians, perhaps with reason, as the driving force of civilization). → 2. **Mais, numerosos historiadores* ***têm considerado ela****, eventualmente com razão, como o verdadeiro motor de civilização.* (More, many historians **have considered her**, perhaps with reason, as the driving force of civilization) 3. *Mais, numerosos historiadores* ***têm-na considerado****, eventualmente com razão, como o verdadeiro motor de civilização.* (More, many historians, **have considered her**, perhaps with reason, as the driving force of civilization).

As mentioned above, when the passive sentence is in the negative form, different steps have to be performed in the transformation to the active sentence. The transformation algorithm takes into account some basic cases, v.g. sentences with *não* (not) or *nunca* (never), but it does not take into account the more complex cases. Example: 1. *Até sábado, Anabela* ***ainda não*** *tinha sido vista por* ***nenhum*** *médico.* (Until Saturday, Anabela had **not yet** been seen by **any** doctor). → 2. **Até sábado,* ***nenhum*** *médico* ***não*** *tinha visto* ***ainda*** *Anabela.* (Until Saturday, **no** doctor had **not yet** seen Anabela). 3. *Até sábado,* ***ainda nenhum*** *médico tinha visto Anabela.* (Until Saturday, **still no** doctor had seen Anabela).

Finally, if a verb is prefixed, and the prefixed form has the same construction as the base form, STRING considers only one lemma and marks, morphologically analyzes the derived prefixed form and associates it with the base lemma adding the information that it features a prefix. Since the algorithm had not taken this information into account, it uses the (non prefixed) lemma to generate the active sentence, thus producing an incorrect (though usually grammatical) paraphrase. In the next example, the verb *inculpar* (indict) has been incorrectly analyzed as a prefixed word, and was associated with the verb *culpar* (blame); the transformation algorithm amplifies the error by using the base verb in the active sentence, instead of using the prefixed word. Notice that sentence 2 is a perfectly grammatical and natural utterance, but not a paraphrase of sentence 1: Example: 1. *O Presidente não pode ser* ***inculpado*** *por um Grande Júri civil.* (The President cannot be indicted by a Grand civil Jury). → 2. *Um Grande Júri civil não pode* ***culpar*** *o Presidente.* (A civil Grand Jury cannot blame the President). 3. *Um Grande Júri civil não pode* ***inculpar*** *o Presidente.* (A Grand civil Jury cannot indict the president).

## VI. Conclusion and Future Work

Exercises on passive transformation are a common feature in language textbooks and online language exercises, both for native and for non-native speakers. This paper addressed the issue of generating active sentences from their passive counterparts, which are extracted from naturally occurring utterances in real texts. The purpose was to create passive-active sentence pairs that will be the basis for the automatic generation (and correction) of exercises on this grammatical topic. Such exercises are to be integrated in the iCALL system REAP.PT [9], [16]. This type of system intends to facilitate the language learning process by providing a large number of automatically generated exercises, which are also automatically corrected (with feedback) so that, on one hand, students can practice autonomously their language skills, and, on the other hand, teachers' time and effort can be used in more creative and relevant tasks.

This iCALL setting raises several NLP and linguistic issues that the paper surveyed. The evaluation and detailed error analysis has shown that perfect results are difficult to obtain, but an important number of situations has been mapped and a course has been set to solve them. Among other linguistic issues, the more relevant to the improvement, not only of the passive-active transformation algorithm, but of the parsing system as a whole, seem to be the following: (i) a more precise identification of passive sentences; (ii) processing of elements in relative sub-clauses prior to other elements in the sentence; (iii) improving the processing of adverbs and adverbial clauses; (iv) improving the processing of negation, not only with negation adverbs but especially with negation determiners; (v) implement a module for processing personal pronouns adequately, taking case and position changes into account; (vi) improving the parsing of coordination, particularly in the case of coordinated `NP` and `PP`; and (vii) improving the processing of punctuation, particularly the use of commas.

This will constitute the focus of future work. Naturally, for each topic above, different strategies may have to be used.

Once results are further improved, this work can be integrated in the REAP.PT iCALL system, as one of its syntactical games. Several types of exercises can be envisioned. A simpler exercise would require the automatic generation of distractors (or foils) for multiple-choice exercises. A more challenging type of exercise would be the implementation of a module that would parse the students' free answer and match it onto the key grammatical features targeted by the exercise. Irrespective of exercise type, the integration will also require the adaptation of the system's graphical interface; the implementation of a help module, a score module and an automatic feedback module.

## Acknowledgment

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

## References

[1] Ait-Mokhtar, S., Chanod, J. & Roux, C. (2002). Robustness beyond shallowness: incremental dependency parsing, *Natural Language Engineering* 8 (2/3): 121–144.

[2] 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.

[3] 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.

[4] 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.

[5] Baptista, J., Costa, N., Guerra, J., Zampieri, M., Cabral, M., and Mamede, N. (2010). P-AWL: Academic Word List for Portuguese. In Proceedings of the 10th International Conference on Computational Processing of the Portuguese Language, *volume 6001 of PROPOR 2010*, pages 120-123. Springer-Verlag, Berlin, Heidelberg.

[6] Baptista, J. Jorge Baptista. ViPEr: A Lexicon-Grammar of European Portuguese Verbs. Proceedings of the *$31^{th}$ International Conference on Lexis and Grammar*, Nové Hrady (Czech Republic), September 19-22, 2012, pp. 10-16.

[7] 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.

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

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

[10] Council of Europe. *Common European Framework of Reference for Languages: learning, teaching, assessment.* Cambridge: Cambridge University Press, 2001.

[11] Figueirinha, P. Syntactic REAP.PT - Exercises on Word Formation. Master's Thesis, Instituto Superior Técnico - Universidade Técnica de Lisboa, 2013.

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

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

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

[15] Martins, R., G. Nunes, & R. Hasegawa. Curupira: A functional parser for Brazilian Portuguese. In N. Mamede, I. Trancoso, J. Baptista, & M. das Graças Volpe Nunes (Eds.), *Computational Processing of the Portuguese Language*, Volume 2721 of *Lecture Notes in Computer Science*, pp. 195-195. Springer Berlin / Heidelberg.

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

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

[18] Raposo, E., Nascimento, F., Mota, M., Segura, L., Mendes, A., Vicente, G. *Gramática do Português*. Fundação Calouste Gulbenkian, 2013.

[19] Silva, A. *Pictorial REAP.PT*. Master's Thesis, Instituto Superior Técnico - Universidade Técnica de Lisboa, 2011.

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