Procesamiento del Lenguaje Natural, Revista nº 47 septiembre de 2011, pp 301-309 recibido 02-05-2011 aceptado 26-05-2011

Ensinador: corpus-based Portuguese grammar exercises

Ensinador: Ejercicios de gramática portuguesa basados en corpus

Alberto Simões
University of Minho
ambs@ilch.uminho.pt

Diana Santos
University of Oslo and FCCN
d.s.m.santos@ilos.uio.no

Resumen: En este artículo describimos una herramienta que crea ejercicios gramaticales para la enseñanza universitaria del portugués como lengua extranjera, utilizando grandes *corpora* anotados. Después de presentar los motivos que nos llevaron a crear dicha herramienta, y el contexto en el que fue desarrollada, presentamos sus especificaciones y sus aplicaciones, analizando algunos requisitos con ejemplos concretos. Presentamos su uso en un contexto de enseñanza real y comparamos esta herramienta con otros sistemas existentes. Terminamos con una breve descripción de trabajos en curso y futuros relacionados con el tema.
Palabras clave: corpora, enseñanza de la lengua asistida por ordenador, corpora anotados, herramientas didácticas, portugués

Abstract: In this paper we describe *Ensinador*, a tool that creates grammar exercises for teaching Portuguese as a foreign language, at university level, based on large annotated corpora. After discussing the motivation for the tool, and the context in which it was deployed, we present its specification and its implementation, discussing the particular requirements with concrete examples. We report actual use of the tool in teaching, and compare it to other systems in the literature, ending the paper with a short description of future and ongoing related work.
Keywords: corpora, computer assisted language learning, annotated corpora, teaching aids, Portuguese

1 Claims on the use of corpora for language teaching

It is a well rehearsed "truth" to say that text corpora are valuable aids for language and linguistics teaching, see e.g. Bacelar do Nascimento (1997). On the other hand, there is no denying that there are many more claims on this than reports of actual use in the classroom, as for example the Teaching and Language Corpora (TaLC) conferences proceedings illustrate.

This stems, in our opinion, from three main reasons:

- the unfortunate separation between corpus compilers and corpus users, who are rarely the same people, as discussed and problematized by Santos (1999);
- individual teaching is — and should be — centered on the individual students and classes, and therefore teachers neither have the urge nor the confidence to report their experiences as generalizable or directly interesting for others;
- despite very honourable exceptions, university and high school language teachers are often the least interested and computer-savvy users of technology and computers. In fact, our experience has been sobering in this respect: After having lectured for more than a decade around Portugal and Brazil on the AC/DC corpora, the two main answers, apart from total neglect, have been: (i) this is not the type of text that interests us/that we base our teaching in; (ii) it is too difficult to use (so they do not even try).

So this paper is a first report on how we decided to change this situation, not as a corpus development community from the outside, but by developing corpus-based teaching aids, from the inside, and using them in the classroom, at the Portuguese grammar courses at the University of Oslo.

This demonstrates on the field the poten-

ISSN 1135-5948

tialities of annotated corpora, and makes the tools user-driven, to an extent that outside lecturers could never achieve.

In what follows, we describe the rationale and the use of *Ensinador*[1] as the first version of a tool suite designed for using the power of the annotated AC/DC corpora to create exercises and teaching materials about Portuguese grammar.

2 Pedagogical assumptions and context

We concur with Sinclair (1997, page 31) that a student should be presented authentic materials from the start.

> [...] *it is almost impossible to invent an adequate example; attempts by language teachers, lexicographers an others to represent usage are often embarrassing and never reliable.*

The same point is also forcefully made by Sardinha (2007) when discussing teaching of English in Brazil.

Moreover, in a time when access to written and spoken Portuguese, thanks to the Internet, is not a problem for any adult student of the language, it is easy to expose or get exposed to language material.

The role of a university teacher, then, is to guide the students in mastering or at least understanding and being aware of the grammatical and cultural idiosyncrasies of the language s/he is teaching, illustrated and motivated by the actual speech of native language speakers.

We note that, in the context of the teaching to adult learners of an international language such as Portuguese, natively spoken in the five continents, there is no point in requiring, or even expecting, a teacher to know intimately all varieties. This speaks even more forcefully for the use of authentic materials in the classroom and outside the classroom.

Besides, no matter the excellency of PLE[2] manuals, most students who attend a university course have contacted directly with Portuguese in a Portuguese speaking country, and have therefore constructed their model of the language, which is richer and more diversified than any manual, which must be written for a general audience, can purport to address.

Finally, we believe that the teacher has some, if not considerable, influence on what the students learn and grasp. Therefore the creation of materials for teaching grammar and culture cannot be done automatically or beforehand a particular curriculum and pedagogical environment is defined and implemented.

Aligning with the current Natural Language Processing (NLP) current who believes in computer-assisted applications (rather than fully automatic ones) — this is a point of view we have strived to defend in Linguateca — we created this program not to replace the teacher in the curriculum design or in the exercise choice, but to aid the deployment of both curriculum and exercises, leaving the teacher in control.

And this is how we offer this service to the whole community of PLE teachers (and possibly also teachers of Portuguese grammar to native speakers): teachers are thus able to create their own materials, by selecting, and even correcting, the material automatically discovered by *Ensinador*.

This is the main difference between our approach and that of VISL (Bick, 2005) and WERTi (Working With English Real Texts) by Meurers et al. (2010), which are meant to be directly used by the learners – with consequent attention to their interface and usability. Of course there are also other differences compared to these older and excellent systems: While WERTi annotates full texts on demand with specific developed grammars, and creates several exercises on top of those texts, it does so for a fixed and preprogrammed number of well-known difficulties for learners of English. Our approach, in contrast, uses pre-annotated (parsed) corpora, but allows any possible kind of context and thus exercise type to be devised by the teacher/user. As to VISL, its setup is currently limited to simpler exercises that can be generalized to the multitude of the languages encompassed by the project, while Ensinador was devised to exploit maximally the idyosincrasies of the Portuguese language (or of the

[1] The name, derived from the verb "ensinar" (to teach), is a pun on "escrevedor" (used in Vargas Llosa's book title in Portuguese *Tia Júlia e o escrevedor*, and is meant to express something feebler and less worthy than a teacher ("professor"), just like "escrevedor" is less than a writer ("escritor"). These subtleties are lost in English, where the corresponding nouns are regularly derived from the verbs.

[2] *Português como língua estrangeira*, that is, Portuguese for foreigners.

annotated corpora behind it). Finally, an important property of both WERti and VISL is their immediate feedback/correction after the user inputs. This is, however, not the way we intend the exercises to be used and understood, as explained later in this paper.

Before we present *Ensinador* in detail, we give some information on the context it was developed, briefly introducing AC/DC and Linguateca.

3 The AC/DC service

One of the main goals of Linguateca, a project devoted to fostering R&D on the computational processing of the Portuguese language (Santos, 2000; Santos, 2009), was to create widely available resources for Portuguese. Therefore, since 1999 (Santos and Bick, 2000) we have been serving access to corpora on the Web as one of the first services on this kind, and this service has progressed and improved ever since, giving as well origin to other more specialized projects and services[3].

Currently (Santos, 2011) the AC/DC service gives access to more than 250 million words from various genres, from more than 20 different corpora, all of them syntactically analysed with PALAVRAS (Bick, 2000) which is a full fledged parser for Portuguese developed in the context of the VISL project by Eckhard Bick. In addition, several other kinds of information are associated with these corpora, and the most relevant for the purposes of the present paper is the semantic information on some domains such as colour, clothing and sentiments[4].

We believe that AC/DC is one of the most powerful and knowledge-rich corpus systems on the Web no matter the language, and several other services have been built on top of it: a frequency service, and a semantic relation discovery service (Freitas et al., 2011) are examples of such tools.

Ensinador is also built on top of AC/DC, and we turn to its functionalities now.

[3]For example, the first treebank for Portuguese, Floresta Sintá(c)tica (Afonso et al., 2002), and parallel corpora such as COMPARA (Frankenberg-Garcia and Santos, 2002) and CorTrad (Tagnin, Teixeira, and Santos, 2009).

[4]Currently only *fear* is available, see (Maia and Santos, 2011).

4 Ensinador: Functionalities

Basically, we wanted to create exercises based on particular features of grammar — and get at the same time the exercise and its solution.

The kind of exercise that *Ensinador* produces is a cloze test, that is, the student is given sentences where one or more words have been removed, and s/he should fill the blanks, therefore restoring the original text. As a very simple example, consider the following sentence to be completed with the correct tense of the verb *ter* (to have):

A minha irmã mais velha morreu quando ____ 88 anos.

Having access to millions of words, the first requirement was to be able to select among the thousands of possible candidates.

In addition, to create a single exercise we might employ different query expressions, and apply them to different corpora[5].

This means that after a sizeable number of selected examples the teacher would like to shuffle them in random order — keeping the exercises and solutions aligned, of course.

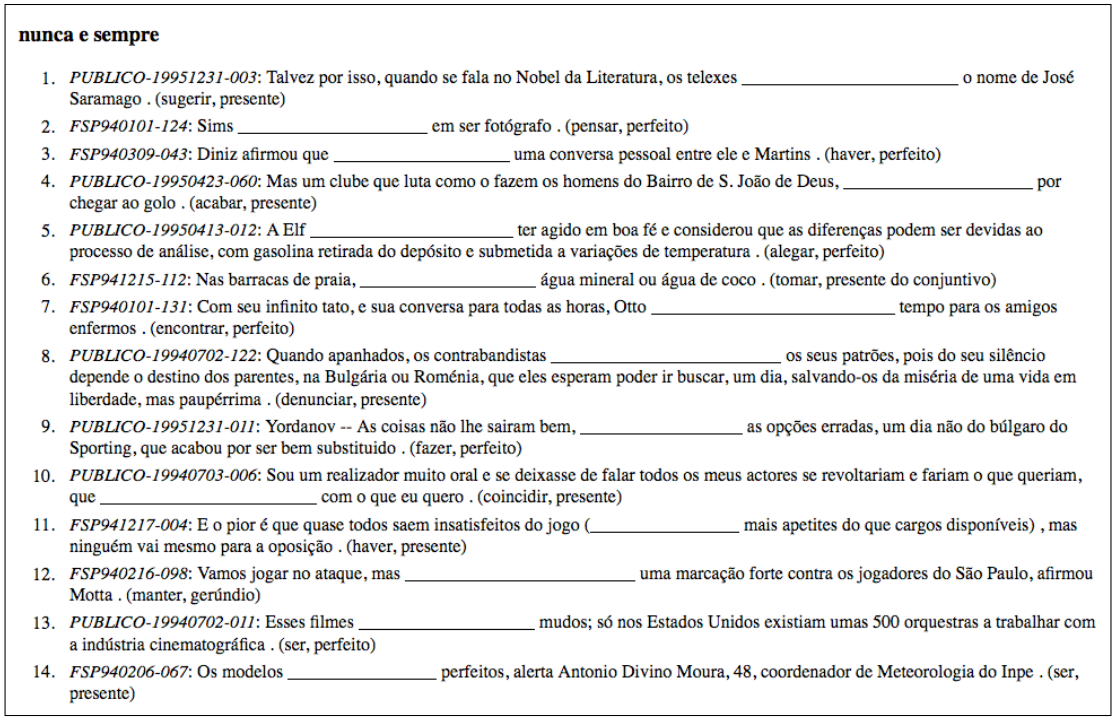
In some cases, when the goal of the exercise concerns for example tense (a morpheme in most cases in Portuguese), in order to recreate the original utterance one would need to know the actual verb that had been removed, so the next requirement was to have it in the lemma form, after the sentence, see figure 1.

But soon other, more complex, cases turned up: for example, to drill the use of quantifiers, it was deemed important to choose the right number of a particular noun[6] and therefore, also the noun lemma was required in a search where more than one word was specified, see figure 2. (In other words, depending on the search expression, and on the actual complexity of the results, different information could be needed to be expressed as an aid to the exercise).

In still other cases, exercises whose goal is to distinguish between false friends or related meanings (such as reflexive vs. non-reflexive

[5]For example, corpora of newspaper text, or oral interviews, or blogs, or literary fiction, or even technical language...

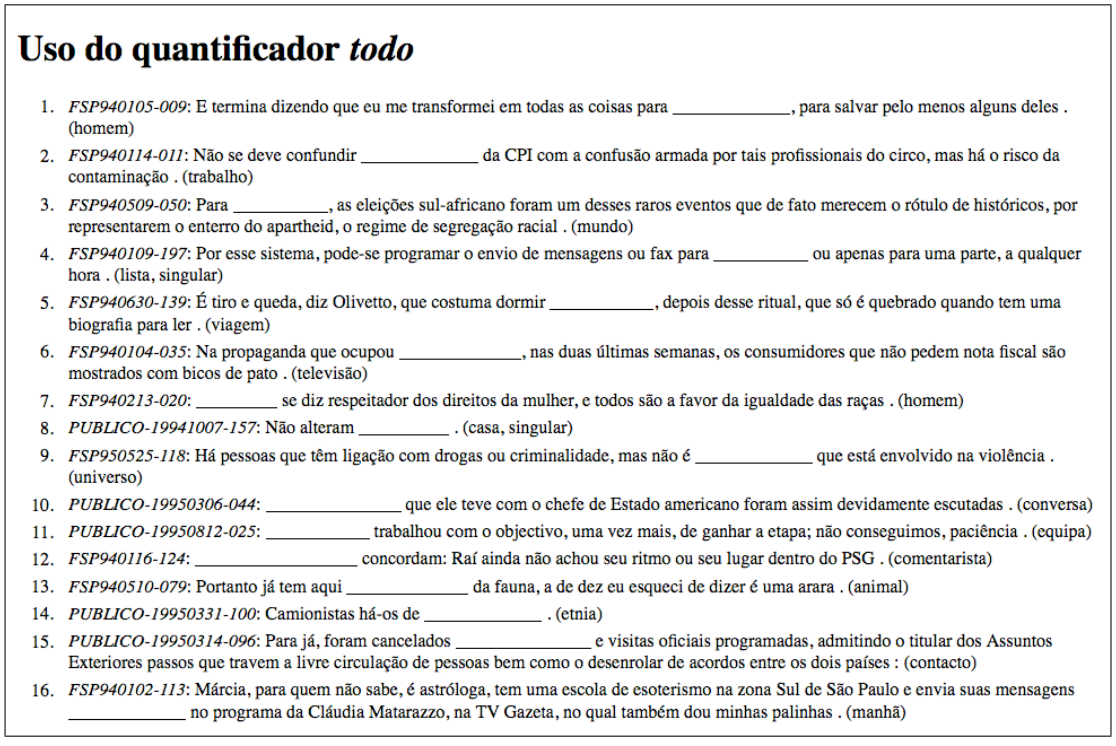
[6]Given that there are the four possibilities *toda a história, todas as histórias, a história toda, as histórias todas* with different meanings and shades, roughly paraphrasable in English by the similar set "each story", "all stories", "every story", "the whole story", "the whole stories".



nunca e sempre

1. *PUBLICO-19951231-003*: Talvez por isso, quando se fala no Nobel da Literatura, os telexes _____ o nome de José Saramago . (sugerir, presente)
2. *FSP940101-124*: Sims _____ em ser fotógrafo . (pensar, perfeito)
3. *FSP940309-043*: Diniz afirmou que _____ uma conversa pessoal entre ele e Martins . (haver, perfeito)
4. *PUBLICO-19950423-060*: Mas um clube que luta como o fazem os homens do Bairro de S. João de Deus, _____ por chegar ao golo . (acabar, presente)
5. *PUBLICO-19950413-012*: A Elf _____ ter agido em boa fé e considerou que as diferenças podem ser devidas ao processo de análise, com gasolina retirada do depósito e submetida a variações de temperatura . (alegar, perfeito)
6. *FSP941215-112*: Nas barracas de praia, _____ água mineral ou água de coco . (tomar, presente do conjuntivo)
7. *FSP940101-131*: Com seu infinito tato, e sua conversa para todas as horas, Otto _____ tempo para os amigos enfermos . (encontrar, perfeito)
8. *PUBLICO-19940702-122*: Quando apanhados, os contrabandistas _____ os seus patrões, pois do seu silêncio depende o destino dos parentes, na Bulgária ou Roménia, que eles esperam poder ir buscar, um dia, salvando-os da miséria de uma vida em liberdade, mas paupérrima . (denunciar, presente)
9. *PUBLICO-19951231-011*: Yordanov -- As coisas não lhe sairam bem, _____ as opções erradas, um dia não do búlgaro do Sporting, que acabou por ser bem substituido . (fazer, perfeito)
10. *PUBLICO-19940703-006*: Sou um realizador muito oral e se deixasse de falar todos os meus actores se revoltariam e fariam o que queriam, que _____ com o que eu quero . (coincidir, presente)
11. *FSP941217-004*: E o pior é que quase todos saem insatisfeitos do jogo (_____ mais apetites do que cargos disponíveis) , mas ninguém vai mesmo para a oposição . (haver, presente)
12. *FSP940216-098*: Vamos jogar no ataque, mas _____ uma marcação forte contra os jogadores do São Paulo, afirmou Motta . (manter, gerúndio)
13. *PUBLICO-19940702-011*: Esses filmes _____ mudos; só nos Estados Unidos existiam umas 500 orquestras a trabalhar com a indústria cinematográfica . (ser, perfeito)
14. *FSP940206-067*: Os modelos _____ perfeitos, alerta Antonio Divino Moura, 48, coordenador de Meteorologia do Inpe . (ser, presente)

Figure 1: Exercise on the position of the *nunca* and *sempre* adverbs.



Uso do quantificador *todo*

1. *FSP940105-009*: E termina dizendo que eu me transformei em todas as coisas para _____, para salvar pelo menos alguns deles . (homem)
2. *FSP940114-011*: Não se deve confundir _____ da CPI com a confusão armada por tais profissionais do circo, mas há o risco da contaminação . (trabalho)
3. *FSP940509-050*: Para _____, as eleições sul-africano foram um desses raros eventos que de fato merecem o rótulo de históricos, por representarem o enterro do apartheid, o regime de segregação racial . (mundo)
4. *FSP940109-197*: Por esse sistema, pode-se programar o envio de mensagens ou fax para _____ ou apenas para uma parte, a qualquer hora . (lista, singular)
5. *FSP940630-139*: É tiro e queda, diz Olivetto, que costuma dormir _____, depois desse ritual, que só é quebrado quando tem uma biografia para ler . (viagem)
6. *FSP940104-035*: Na propaganda que ocupou _____, nas duas últimas semanas, os consumidores que não pedem nota fiscal são mostrados com bicos de pato . (televisão)
7. *FSP940213-020*: _____ se diz respeitador dos direitos da mulher, e todos são a favor da igualdade das raças . (homem)
8. *PUBLICO-19941007-157*: Não alteram _____ . (casa, singular)
9. *FSP950525-118*: Há pessoas que têm ligação com drogas ou criminalidade, mas não é _____ que está envolvido na violência . (universo)
10. *PUBLICO-19950306-044*: _____ que ele teve com o chefe de Estado americano foram assim devidamente escutadas . (conversa)
11. *PUBLICO-19950812-025*: _____ trabalhou com o objectivo, uma vez mais, de ganhar a etapa; não conseguimos, paciência . (equipa)
12. *FSP940116-124*: _____ concordam: Raí ainda não achou seu ritmo ou seu lugar dentro do PSG . (comentarista)
13. *FSP940510-079*: Portanto já tem aqui _____ da fauna, a de dez eu esqueci de dizer é uma arara . (animal)
14. *PUBLICO-19950331-100*: Camionistas há-os de _____ . (etnia)
15. *PUBLICO-19950314-096*: Para já, foram cancelados _____ e visitas oficiais programadas, admitindo o titular dos Assuntos Exteriores passos que travem a livre circulação de pessoas bem como o desenrolar de acordos entre os dois países : (contacto)
16. *FSP940102-113*: Márcia, para quem não sabe, é astróloga, tem uma escola de esoterismo na zona Sul de São Paulo e envia suas mensagens _____ no programa da Cláudia Matarazzo, na TV Gazeta, no qual também dou minhas palinhas . (manhã)

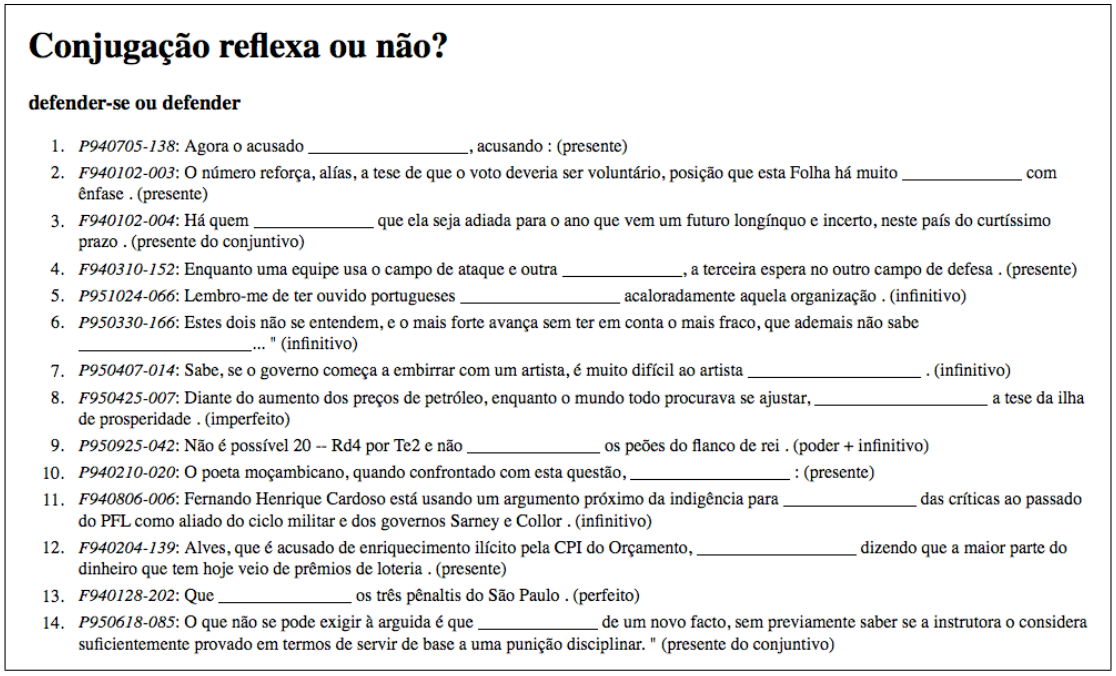
Figure 2: Exercise on the *todo* quantifier (position and form)

conjugation) require that the student is given the verbal tense and person so that s/he can insert the right lexeme in the correct form; see figure 3.

The result is then obtained as simple HTML pages (one for the exercises and other for the solution), which can easily be edited by the teacher afterwards.

5 *Ensinador: how it works*

Ensinador is a web application. Its basic behavior is similar to a simple concordance sys-



Conjugação reflexa ou não?

defender-se ou defender

1. *P940705-138*: Agora o acusado _____, acusando : (presente)
2. *F940102-003*: O número reforça, alías, a tese de que o voto deveria ser voluntário, posição que esta Folha há muito _____ com ênfase . (presente)
3. *F940102-004*: Há quem _____ que ela seja adiada para o ano que vem um futuro longínquo e incerto, neste país do curtíssimo prazo . (presente do conjuntivo)
4. *F940310-152*: Enquanto uma equipe usa o campo de ataque e outra _____, a terceira espera no outro campo de defesa . (presente)
5. *P951024-066*: Lembro-me de ter ouvido portugueses _____ acaloradamente aquela organização . (infinitivo)
6. *P950330-166*: Estes dois não se entendem, e o mais forte avança sem ter em conta o mais fraco, que ademais não sabe _____... " (infinitivo)
7. *P950407-014*: Sabe, se o governo começa a embirrar com um artista, é muito difícil ao artista _____ . (infinitivo)
8. *F950425-007*: Diante do aumento dos preços de petróleo, enquanto o mundo todo procurava se ajustar, _____ a tese da ilha de prosperidade . (imperfeito)
9. *P950925-042*: Não é possível 20 -- Rd4 por Te2 e não _____ os peões do flanco de rei . (poder + infinitivo)
10. *P940210-020*: O poeta moçambicano, quando confrontado com esta questão, _____ : (presente)
11. *F940806-006*: Fernando Henrique Cardoso está usando um argumento próximo da indigência para _____ das críticas ao passado do PFL como aliado do ciclo militar e dos governos Sarney e Collor . (infinitivo)
12. *F940204-139*: Alves, que é acusado de enriquecimento ilícito pela CPI do Orçamento, _____ dizendo que a maior parte do dinheiro que tem hoje veio de prêmios de loteria . (presente)
13. *F940128-202*: Que _____ os três pênaltis do São Paulo . (perfeito)
14. *P950618-085*: O que não se pode exigir à arguida é que _____ de um novo facto, sem previamente saber se a instrutora o considera suficientemente provado em termos de servir de base a uma punição disciplinar. " (presente do conjuntivo)

Figure 3: Exercise on the choice between reflexive or non-reflexive conjugation

tem: it requires the user to introduce a search string, using the Corpus Query Processor (CQP) syntax (Christ et al., 1999; Evert and the OCWB Development Team, 2010), and to select the corpus to be searched. Optionally, the user might choose to define a title for the exercise section being created. Figure 4 shows this interface.

Ensinador

A procurar no corpo "CHAVE"

Procurar para a secção

Procurar

Corpo "CHAVE"

Sobre:

O corpus **CHAVE** contém textos jornalísticos dos jornais Público e Folha de São Paulo dos anos de 1994 e 1995. Este corpus foi criado no âmbito da participação da língua portuguesa no CLEF.

Tipo	Jornalístico
Variante(s)	PT BR
Tamanho (unidades)	123.9 milhões
Tamanho (palavras)	99.4 milhões
Anotado	não

Marcadores estruturais: **DOC** [documento], s [frase], **CATEGORY** (secção do jornal), **AUTHOR** (autor do texto), **NE** (lista de nomes próprios contidos no documento, usando a anotação do PALAVRAS).

Veja um excerto do corpo e informação adicional.

Agradecimentos: Agradecemos ao PÚBLICO e à Folha de São Paulo a gentil autorização de disponibilização dos textos.

Procure noutros corpos: AmostRA-NILC | ANCIB | Avante! | CD HAREM | CETEMPúblico | CHAVE | CONDIVport | CoNE | DiaCLAV | ECI-EBR | ECI-EE | ENPCPUB (parte em português) | Experiencias | Floresta | FrasesPB | FrasesPP | Museu da Pessoa | Natura/Minho | NILC/São Carlos | Vercial |

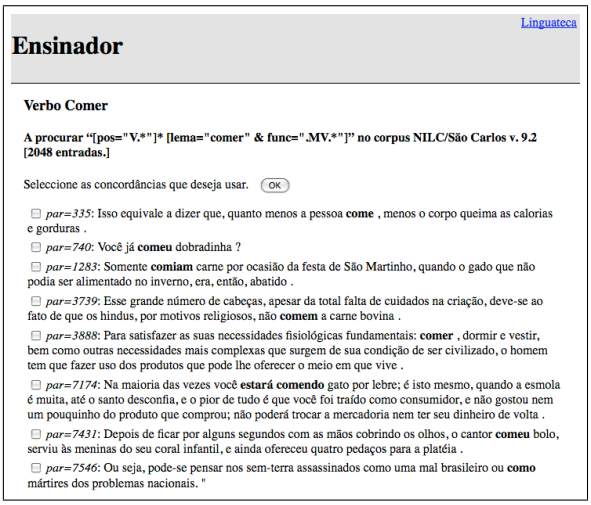
Última actualização: 22 de Março de 2011

Perguntas, comentários e sugestões

Figure 4: Start point when using Ensinador.

Although any well formed query can be used, it is important to note that it is the hits that are replaced by blanks. It is therefore possible, and useful, to specify cases where both simple and complex tenses (for example) may be used. (The same space, however, should be provided in the exercise, not to uncover the solution.)

For example, one might search by the verb "comer" in any of its forms, using "`[pos="V.*"]* [lema="comer" & func=".MV.*"]`." The tool will present a page with the results in concordance form. An example of searching for this expression in a test corpus is presented in figure 5.



Linguateca

Ensinador

Verbo Comer

A procurar "[pos="V.*"]* [lema="comer" & func=".MV.*"]" no corpus NILC/São Carlos v. 9.2 [2048 entradas.]

Seleccione as concordâncias que deseja usar. OK

- *par=335*: Isso equivale a dizer que, quanto menos a pessoa **come** , menos o corpo queima as calorias e gorduras .
- *par=740*: Você já **comeu** dobradinha ?
- *par=1283*: Somente **comiam** carne por ocasião da festa de São Martinho, quando o gado que não podia ser alimentado no inverno, era, então, abatido .
- *par=3739*: Esse grande número de cabeças, apesar da total falta de cuidados na criação, deve-se ao fato de que os hindus, por motivos religiosos, não **comem** a carne bovina .
- *par=3888*: Para satisfazer as suas necessidades fisiológicas fundamentais: **comer** , dormir e vestir, bem como outras necessidades mais complexas que surgem de sua condição de ser civilizado, o homem tem que fazer uso dos produtos que pode lhe oferecer o meio em que vive .
- *par=7174*: Na maioria das vezes você **estará comendo** gato por lebre; é isto mesmo, quando a esmola é muita, até o santo desconfia, e o pior de tudo é que você foi traído como consumidor, e não gostou nem um pouquinho do produto que comprou; não poderá trocar a mercadoria nem ter seu dinheiro de volta .
- *par=7431*: Depois de ficar por alguns segundos com as mãos cobrindo os olhos, o cantor **comeu** bolo, serviu às meninas do seu coral infantil, e ainda ofereceu quatro pedaços para a platéia .
- *par=7546*: Ou seja, pode-se pensar nos sem-terra assassinados como uma mal brasileiro ou **como** mártires dos problemas nacionais. "

Figure 5: Ensinador presenting concordances to the user.

Note that each concordance has a check box at the left. They are used by the teacher to choose the example sentences for the exercise, according to his or her personal taste and appropriateness to the grammatical points s/he wishes to illustrate. After choosing the sentences to be used and clicking the "OK" button, *Ensinador* presents the current status of the exercise, as shown in figure 6.

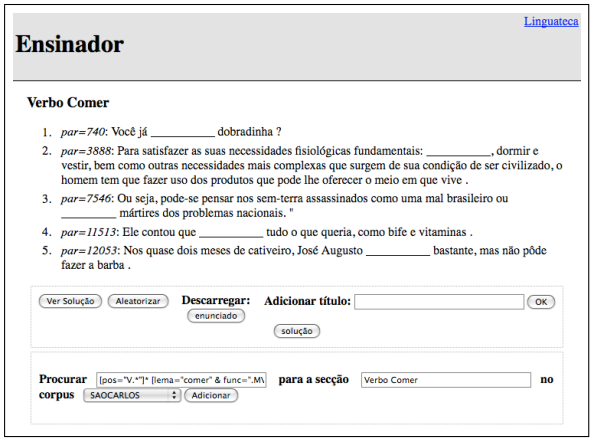


Figure 6: Ensinador presenting status of the exercise.

In this step the tool shows the exercise as it would be presented to the student, with the part of the sentence that matched with the query replaced by a blank space to be filled in by the student.

At this point the teacher is able to do several tasks, such as

- to see the solution, that is, this same interface but with the blanks replaced by the respective words;
- to scramble the order of the sentences (as many times as s/he wishes);
- to download the exercise or the solution in HTML format (without the button controls);
- to add a global title to the exercise.

Also, it is possible to perform new searches, and add them to the current exercise. This new search can use a different search expression, can add the lines to the same exercise section (under the same title) or in a new one, as well as change the corpus where the expression will be searched.

As discussed previously, to produce meaningful exercises it is often necessary to provide information about what should fill the blanks. To allow for this, the CQP syntax was extended in order to let the teacher select extra information to be presented to the student. This extension allows the teacher to add one or more attributes (provided they are encoded in the corpus) to be added to the end of the sentence, in parenthesis. For example, the query "`[pos="V"].lema`" will find all sentences with verbs, and present the verb lemma:

O atendimento que a casa ____ às mulheres é através de reuniões e cursos. (fazer)

Another example: In order to request verbs with clitics, one could employ a query such as `[pos="PRON"].pessnum [pos="V"].lema.temcagr`, which would then produce exercises like the following:

Quem _____ dinheiro era imediatamente participado à polícia. (ela, dar, imperfeito do conjuntivo)

6 Implementation

Ensinador is built over the AC/DC corpora, that are encoded in the Open IMS Corpus Workbench (Evert and the OCWB Development Team, 2010), and tagged with PALAVRAS (Bick, 2000). The application interface is written in the Perl programming language, using the `CWB::CQP` module, and interfaces with PHP configuration files using the `PHP::Include` module.

The tool is built using a Common Gateway Interface (CGI), using a minimal amount of Java Script that is controlled by jQuery[7]. It does not use any kind of session or cookie management, making it easy to use with any existent browser.

The most relevant part in the overall algorithm is how the CQP language was modified, and how the extra properties of words are computed: A partial parser of the default CQP language was implemented, and extended to interpret one or more optional attributes after each token. As this is an extension, it is not possible to easily detect repeating tokens, and therefore the attributes can only be used in non-repeating or non-optional query expression parts. For example, one could not add the name of an attribute to the two first units of the next query, only to the first and to the last, in "`[pos="V"] []+ [pos="N"].lema`", as Ensinador is not able to know how many tokens were matched in the repeating expression. When multiple optional or repeating tokens are present, only the fixed tokens in the beginning or the end of the expression can be annotated.

After associating each attribute with the respective token position, the tool activates and deactivates attributes, one at a time, to fetch the information to be presented be-

[7] Available from `http://jquery.com/`.

tween parentheses. This can be a very inefficient procedure, so Ensinador only computes this information once for each concordance, caching subsequently the information.

7 Use in Teaching

Even though this project is quite young, we have used *Ensinador*'s results while teaching Portuguese grammar at the Department of Literature, Area Studies and Languages (ILOS) at the Arts Faculty of the University of Oslo in the 2011 spring term, for students of beginning and intermediate levels in Portuguese grammar (all of them already proficient in Portuguese, to various degrees).

The way this was done was as follows: after the theoretical material had been presented, the students would try to do the first part of the exercises in class. Doubts about the exercise, as well as comments about particular cases or difficulties were discussed in class, before the students took it home to complete. Then, after a week, the results (i.e., the original renderings) were made available to the students, who could compare them with their own work. Students were encouraged to come back to the teacher if they did not understand why a particular rendering had been chosen by its author, but they never did.

One should note that it was consistently emphasized by the teacher that in some cases the "solution" was merely a question of stylistic choice by the text author; in other cases, one would need to access a larger context to decide; while in some cases only one possibility was grammatically correct.

In a way, this should have been highlighted later when the solutions were made available, but was not (yet) done. We plan to do this new analysis in the Fall semester, so that students can consolidate their grammar skills by reflecting upon their choices and the authors'. They will also be encouraged to express the meaning differences when more than one solution is possible.

All in all, the students' exposition to several kinds of text genres (and varieties) by reading these authentic sentences was generally felt to be positive. In fact, an added advantage was that it was clear which variety (from Portugal or Brazil) the sentences belonged to, thanks to the codes associated with the concordance lines. This allowed a better understanding of the commonalities and specificities of the two varieties in the grammatical subjects at hand.

8 Further Work

There are two distinct schools in foreign language teaching as far as the use of translation materials in foreign language teaching is concerned (Tavares, 2008), but we won't argue for any of these positions here.

Rather, we note that, no matter the position as far as **language** teaching is concerned, looking at translation practice is absolutely required when you teach **translation**, and one can reuse a variation of the present tools to teach translation if one has access to aligned parallel corpora encoded in a AC/DC-like format like COMPARA or CorTrad, or the ones being developed in PerFide (Araújo et al., 2010).

One can then present translation pairs with the source or the target language items removed, for the constructions or lexical items that the teacher is interested in, as illustrated by Frankenberg-Garcia (1999).

In order to concentrate precisely on the issues that have a non-standard solution, or for which several genuinely different translation solutions have been put forward, it would be advantageous to have a corpus of many translations on which to base the *Trans-ensinador*. As pointed out by Malmkjaer (1996), the points where there is disagreement or multiple solutions are those which illuminate best the differences between the languages and the complexities of the translation among them.

We plan thus in the near future to deploy *Trans-ensinador*, based on parallel corpora. In order to offer a richer set of options, we plan to use a word-aligner, so that the query expression can be more precise and the translation part better highlighted than if simple pairs of (aligned) sentences were presented. This will be done with NATools (Simões and Almeida, 2003) using an approach already applied to COMPARA (Santos and Simões, 2008), but which needs to be extended so that it can be used in a search expression as well.

Other AC/DC based teaching tools have also been deployed but have so far not being used in actual teaching, namely:

- a comparison tool that allows the user to contrast two different queries;
- a quantitative distribution tool that

gives a bird's eye of the frequency of several phenomena.

- the possibility to search by semantic relation, that is, all cases that have *animal* as hypernym, or are antonyms of *bom* ("good"), or synonyms of *mesa* ("table"), by annotating the AC/DC corpora with this additional information.

9 Concluding remarks

With this paper, we presented a working NLP application in CALL (or CALT, computer-aided language teaching) which radically increases a teacher's power while using resources compiled for the computational processing of Portuguese. The system is available from `http://www.linguateca.pt/Ensinador/` and its open source code can be downloaded by whoever wants to use it in connection with their own corpora (possibly in other languages). We should anyway stress that the wealth of the corpus material and underlying annotation is the result of a long-term commitment by Linguateca to making resources available for Portuguese and maintaining them in a long term perspective.

Acknowledgments

The work described here was developed in the scope of Linguateca, which has been funded by the Portuguese government, EU (FEDER and FSE) `POSC/339/1.3/C/NAC`, UMIC and FCCN. The work in Trans-Ensinador is being also funded by the project Per-fide, *Português em paralelo com seis línguas (Português, Español, Russian, Français, Italiano, Deutsch, English)* grant `PTDC/CLE-LLI/108948/2008` from *Fundação para a Ciência e a Tecnologia*.

References

Afonso, Susana, Eckhard Bick, Renato Haber, and Diana Santos. 2002. Floresta sintá(c)tica: um treebank para o português. In Anabela Gonçalves and Clara Nunes Correia, editors, *Actas do XVII Encontro Nacional da Associação Portuguesa de Linguística (APL 2001)*, pages 533–545, Lisboa, Portugal, 2-4 de Outubro de 2001. APL.

Araújo, Sílvia, José João Almeida, Alberto Simões, and Idalete Dias. 2010. Apresentação do projecto Per-Fide: Paralelizando o português com seis outras línguas. *Linguamática*, 2(2):71–74, June.

Bacelar do Nascimento, Maria Fernanda. 1997. A exploração de corpora linguísticos no ensino/aprendizagem do português. In *Seminário Internacional do Português como Língua Estrangeira*, Macau, 21–24, Maio.

Bick, Eckhard. 2000. *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Ph.D. thesis, Aarhus University.

Bick, Eckhard. 2005. Grammar for fun: IT-based grammar learning with VISL. In P. Juel, editor, *CALL for the Nordic Languages*, pages 49–64, Copenhagen. Samfundslitteratur.

Christ, Oliver, Bruno M. Schulze, Anja Hofmann, and Esther König, 1999. *The IMS Corpus Workbench: Corpus Query Processor (CQP): User's Manual*. Institute for Natural Language Processing, University of Stuttgart, March.

Evert, Stefan and the OCWB Development Team. 2010. The ims open corpus workbench (cwb) cqp query language tutorial (cwb version 3.0), 17 February. `http://cwb.sourceforge.net/files/CQP_Tutorial.pdf`.

Frankenberg-Garcia, Ana. 1999. Crosslinguistic influence as a key to extracting second language teaching materials for monolingual classes from translation corpora. In Sylviane Granger, editor, *Proceedings of the Workshop Contrastive Linguistics and Translation Studies: Empirical Approaches*, 5-6 February.

Frankenberg-Garcia, Ana and Diana Santos. 2002. COMPARA, um corpus paralelo de português e de inglês na Web. *Cadernos de Tradução*, IX(1):61–79.

Freitas, Cláudia, Diana Santos, Hugo Gonçalo Oliveira, and Violeta Quental. 2011. VARRA: Validação, Avaliação e Revisão de Relações semânticas no AC/DC. In *Atas do ELC2010*.

Maia, Belinda and Diana Santos. 2011. Who us afraid of ... what? Fear in English and Portuguese. In *32th ICAME*.

Malmkjaer, Kirsten. 1996. Who walked in the emperor's garden: The translation of

pronouns in Hans Christian Andersens introductory passages. In Gunilla Anderman and C. Banér, editors, *Proceedings of the Tenth Biennial Conference of the British Association of Scandinavian Studies*. The University of Surrey, Department of Linguistics and International Studies.

Meurers, Detmar, Ramon Ziai, Luiz Amaral, Adriane Boyd, Aleksandar Dimitrov, Vanessa Metcalf, and Niels Ott. 2010. Enhancing authentic web pages for language learners. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 10–18, Stroudsburg, PA, USA, June. Association for Computational Linguistics.

Santos, Diana. 1999. Disponibilização de corpora de texto através da WWW. In Palmira Marrafa and Maria Antónia Mota, editors, *Linguística Computacional: Investigação Fundamental e Aplicações. Actas do I Workshop sobre Linguística Computacional da Associação Portuguesa de Linguística (Lisboa, 25-27 de Maio de 1999)*, pages 323–335, Lisboa. Colibri.

Santos, Diana. 2000. O projecto Processamento Computacional do Português: Balanço e perspectivas. In Maria das Graças Volpe Nunes, editor, *V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR 2000)*, pages 105–113, São Paulo, 19-22 de Novembro. ICMC/USP.

Santos, Diana. 2009. Caminhos percorridos no mapa da portuguesificação: A Linguateca em perspectiva. *Linguamática*, 1(1):25–59, Maio.

Santos, Diana. 2011. Linguateca's infrastructure for Portuguese and how it allows the detailed study of language varieties. *OSLa: Oslo Studies in Language*, 3.

Santos, Diana and Eckhard Bick. 2000. Providing internet access to portuguese corpora: the AC/DC project. In Maria Gavrilidou et al, editor, *Second International Conference on Language Resources and Evaluation, LREC 2000*, pages 205–210, Athens, May–June.

Santos, Diana and Alberto Simões. 2008. Portuguese-English word alignment: some experiments. In *LREC 2008 — The 6th edition of the Language Resources and Evaluation Conference*, Marrakech, 28–30, May. European Language Resources Association (ELRA).

Sardinha, Tony Berber. 2007. The book is NOT on the table: Autenticidade e idiomaticidade do texto para o ensino de inglês na perspectiva da lingüística de corpus. In Maria Cristina Damianovic, editor, *Material Didático: Elaboração e Avaliação*. Taubaté: Cabral Editora, pages 273–286.

Simões, Alberto M. and J. João Almeida. 2003. NATools – a statistical word aligner workbench. *Procesamiento del Lenguaje Natural*, 31:217–224, September.

Sinclair, John. 1997. Corpus evidence in language description. In Anne Wichmann, Steven Fligelstone, Tony McEnery, and Gerry Knowles, editors, *Teaching and language corpora*. Longman, London & New York, pages 27–39.

Tagnin, Stella E.O., Elisa Duarte Teixeira, and Diana Santos. 2009. CorTrad: a multiversion translation corpus for the Portuguese-English pair. *Arena Romanistica*, 4:314–323.

Tavares, Ana. 2008. *Ensino / Aprendizagem do Português como Língua Estrangeira: Manuais de iniciação*. Lidel, Lisboa.