



UNIVERSIDADE DE
COIMBRA



Cristina Mamede Abrantes

*INVESTIGAÇÃO EM CORPORA INFORMATIZADOS DE PRODUÇÕES
ORAIS E ESCRITAS DE APRENDENTES DE PORTUGUÊS
LÍNGUA NÃO MATERNA – FAQ E ORIENTAÇÕES
PARA A EXPLORAÇÃO DE VALÊNCIAS*

*Projeto de Mestrado em Português Língua Estrangeira e Língua Segunda,
orientado pela Professora Doutora Cristina dos Santos Martins e pela Professora
Doutora Isabel Maria de Almeida Santos, apresentado ao Departamento de
Línguas, Literaturas e Culturas da Faculdade de Letras da Universidade de
Coimbra.*

Setembro de 2019

FACULDADE DE LETRAS

INVESTIGAÇÃO EM *CORPORA* INFORMATIZADOS DE PRODUÇÕES ORAIS E ESCRITAS DE APRENDENTES DE PORTUGUÊS LÍNGUA NÃO MATERNA - FAQ E ORIENTAÇÕES PARA A EXPLORAÇÃO DE VALÊNCIAS

Ficha Técnica

Tipo de trabalho	Projeto
Título	Investigação em <i>corpora</i> informatizados de produções orais e escritas de aprendentes de português língua não materna - FAQ e orientações para a exploração de valências
Autor/a	Cristina Mamede Abrantes
Orientadoras	Cristina dos Santos Martins Isabel Maria de Almeida Santos
Júri	Presidente: Doutora Maria Isabel Pires Pereira Vogais: 1. Doutora Tânia Santos Ferreira 2. Doutora Isabel Maria de Almeida Santos
Identificação do Curso	2º Ciclo em Português como Língua Estrangeira e Língua Segunda
Área científica	Linguística
Data da defesa	09-10-2019
Classificação	18 valores



UNIVERSIDADE D
COIMBRA



À Letícia e à Júlia,

sempre.

RESUMO

A definição de um conjunto de perguntas (FAQ) para exploração de valências de pesquisa em *corpora* informatizados de produções escritas e orais de aprendentes de português língua não materna, a partir de uma plataforma digital, tem como principal objetivo orientar o utilizador na extração de dados consistentes destes acervos.

Na génese deste projeto estão dois *corpora* que disponibilizam dados de aprendentes de português língua não materna com vista à sustentação empírica da investigação no âmbito da aquisição/aprendizagem de língua não materna, à fundamentação de práticas pedagógicas e, ainda, à criação de novos instrumentos didáticos: o *corpus* de Produções Escritas de Aprendentes de PL2 (PEAPL2_PLE) e o *Corpus* Oral de Português L2 (COraL-Co).

Estes *corpora* foram disponibilizados *online* a partir do TEITOK, uma plataforma que possibilita a pesquisa através do *Corpus Query Processor* (CQP), uma ferramenta flexível e eficiente, mas cuja rentabilização está condicionada, essencialmente, pelos processos de inserção e edição dos dados na plataforma e pela complexidade das expressões de pesquisa.

Assim, as virtualidades da pesquisa em *corpora* informatizados de aprendentes, por um lado, e os constrangimentos à pesquisa que se podem colocar, por outro, motivaram a definição de um conjunto de FAQ com o objetivo de orientar o utilizador na exploração de *corpora* e rentabilizar as potencialidades da plataforma de pesquisa. As perguntas e as respetivas respostas encontram-se disponíveis para consulta *online*, na página de pesquisa de cada um dos referidos *corpora*, em dois formatos (*PDF* e *mp4*).

Este modelo de construção de FAQ, baseado nos objetivos inerentes à construção de *corpora* e nas valências de pesquisa do CQP, potencia a atualização e o aperfeiçoamento deste tipo de *help search* e a sua replicação em *corpora* desenvolvidos a partir do TEITOK.

Palavras-chave: *Corpora* informatizados, língua não materna, PEAPL2, COraL-Co, Teitok, FAQ.

ABSTRACT

The definition of a set of questions (FAQ) for the exploration of research valences in computerised corpora of oral and written production of Portuguese as a Second or Foreign Language learners, on a digital platform, has as main goal to guide the user in the extraction of data from such collection.

In the genesis of this project are two corpora that provide data of Portuguese as a Second or Foreign Language learners for the empirical support of the research in the scope of the acquisition/learning of a second or foreign language, the rationale of education practices, and the creation of didactic instruments: the *Corpus* de Produções Escritas de Aprendentes de PL2 (PEAPL2_PLE) and the *Corpus* Oral de Português L2 - Coimbra (COral-Co).

These *corpora* have been made available online on TEITOK, a platform that allowed the research through the *Corpus Query Processor* (CQP), a flexible and efficient tool, but whose effectiveness is conditioned, mainly, by the data insertion and editing processes in the platform, and by the complexity of the research expressions.

As such, the virtualities of the research in computerised *corpora* of learners, on one hand, and the restraints that may occur on the research, on the other, have led to the definition of a set of FAQ with the purpose of guiding the user in the exploration of *corpora* and retrieving higher benefit from the research platform capabilities. The questions and the corresponding answers are available for online check in the research page of each of the mentioned *corpora*, in two formats (*PDF* and *mp4*).

This model of FAQ construction, based on the goals inherent to the construction of *corpora* and the CQP's research valences, empower the update and perfecting of this type of *help search* and its replication in *corpora* developed in TEITOK.

Keywords: Computerised *Corpora*, Second and Foreign Language, PEAPL2, COral-Co, Teitok, FAQ

ÍNDICE

Resumo	iv
Abstract	v
Índice de tabelas	ix
Lista de abreviaturas	x
INTRODUÇÃO	1
PARTE I - ENQUADRAMENTO	3
INTRODUÇÃO	3
CAPÍTULO 1 - CORPUS LINGUÍSTICO E CORPUS DE APRENDENTES DE L2	5
INTRODUÇÃO	5
1.1. A modernidade do conceito de <i>corpus</i> : do acervo de dados ao <i>corpus</i> anotado	7
1.2. <i>Corpora</i> de aprendentes de L2	16
1.2.1. Considerações prévias: aquisição e aprendizagem de L2	16
1.2.2. Conceito	18
1.2.3. Virtualidades dos <i>corpora</i> de aprendentes de L2	24
CAPÍTULO 2 - TEITOK: UMA FERRAMENTA PARA ARMAZENAMENTO, PREPARAÇÃO E PESQUISA DE DADOS	26
INTRODUÇÃO	26
2.1. A criação de um <i>corpus</i> informatizado	29
2.2. A preparação e edição de dados	31
2.3. Pesquisa	39
CAPÍTULO 3 - DOIS CORPORA DE APRENDENTES DE PL2	48
INTRODUÇÃO	48
3.1. <i>Corpus</i> de Produções Escritas de Aprendentes de PL2: subcorpus de Português Língua Estrangeira (PEAPL2_PLE)	49
3.2. <i>Corpus</i> Oral de Português L2 (COraL-Co)	56

PARTE II - O PROJETO	63
INTRODUÇÃO	63
CAPÍTULO 1 - APRESENTAÇÃO DO PROJETO E DA METODOLOGIA	64
INTRODUÇÃO	64
1.1. Apresentação do projeto	65
1.2. Metodologia	67
CAPÍTULO 2 - RESULTADOS - FAQ E RESPOSTAS	71
INTRODUÇÃO	71
2.1. Fundamentos para a elaboração das FAQ	72
a. Áreas de investigação no âmbito do PLNM	72
b. Domínios de pesquisa	74
i. Anotação do <i>corpus</i>	74
ii. Construção da pesquisa	74
iii. Armazenamento de dados	76
2.2. Apresentação das FAQ	78
2.3. Construção das respostas e disponibilização na plataforma	82
Considerações finais sobre as FAQ na orientação de pesquisa de <i>corpora</i> de aprendentes	86
BIBLIOGRAFIA	88

ANEXOS	91
ANEXO I - SISTEMA DE ETIQUETAS MORFOSSINTÁTICAS UTILIZADAS NO TEITOK	92
ANEXO II - FAQ E RESPOSTAS (PDF) DISPONÍVEIS NA PÁGINA DE PESQUISA DO PEAPL2_PLE ...	97
ANEXO III - FAQ E RESPOSTAS (PDF) DISPONÍVEIS NA PÁGINA DE PESQUISA DO Coral-Co	139

ÍNDICE DE TABELAS

Tabela 1: Convenções de transcrição utilizadas no PEAPL2_PLE, com base em Leiria (2006).	50
Tabela 2: Convenções adotadas, no COral-Co, na transcrição de produções orais.	59
Tabela 3: Áreas de investigação no âmbito do PLNM (com base na consulta de <i>“Bibliografia sobre aquisição, aprendizagem e ensino do Português Europeu como Língua não Materna”</i>).	72/73
Tabela 4: Conjunto de 18 FAQ, distribuídas por domínios de pesquisa, referentes à pesquisa no PEAPL2_PLE.	78/79
Tabela 5: Conjunto de 18 FAQ, distribuídas por domínios de pesquisa, referentes à pesquisa no COral-Co.	79/80
Tabela 6: FAQ distribuídas por áreas de investigação  e por domínio de pesquisa 	80/81

LISTA DE ABREVIATURAS

COral-Co - *Corpus Oral de Português L2*

CQP - *Corpus Query Processor*

FAQ - *Frequently Asked Question(s)*

HTML - *Hypertext Markup Language*

L1 - Língua materna

L2 - Língua estrangeira/segunda

LA - Língua-alvo

LNМ - Língua Não Materna

mp3 - especificação de ficheiro áudio

mp4 - especificação de ficheiro áudio e vídeo

PDF - *Portable Document Format*

PEAPL2 - *Corpus de Produções Escritas de Aprendentes de PL2*

PEAPL2_PLE - subcorpus de Português Língua Estrangeira

PL2 - Português Língua Estrangeira

PLNM - Português Língua Não Materna

QECRL - *Quadro Europeu Comum de Referência para as Línguas* (Conselho da Europa 2001)

TEI - *Text Encoding Initiative*

XML - *Extensible Markup Language*

INTRODUÇÃO

O presente projeto assenta na definição de um conjunto de perguntas (FAQ) para exploração de valências de pesquisa em *corpora* informatizados de produções escritas e orais de aprendentes de português língua não materna a partir de uma plataforma digital.

A pesquisa no *corpus* de Produções Escritas de Aprendentes de PL2 (PEAPL2) e no *Corpus* Oral de Português L2 (COral-Co) centra-se em textos (escritos e orais) produzidos por aprendentes de português língua não materna e permite extrair dados empíricos com vista à sustentação da investigação no âmbito da aquisição/aprendizagem de língua não materna, à fundamentação de práticas pedagógicas e, ainda, à criação de novos instrumentos didáticos.

O TEITOK, criado por Maarten Janssen (2014), é uma ferramenta informática que permite realizar pesquisas em *corpora* de aprendentes através do *Corpus Query Processor* (CQP), uma ferramenta de pesquisa bastante flexível e eficiente, mas cuja rentabilização está condicionada pelos processos prévios de criação e de edição de *corpora* e por uma linguagem de pesquisa, por vezes, complexa e pouco intuitiva.

As virtualidades da pesquisa em *corpora* informatizados de aprendentes, por um lado, e os constrangimentos à pesquisa que se podem colocar, por outro, foram o ponto de partida para a definição de um conjunto de FAQ, para consulta *online* na área de pesquisa, cujo objetivo principal é orientar o utilizador na extração de dados, de forma consistente, destes acervos. As perguntas organizam-se, assim, em três níveis distintos - anotação do *corpus*, pesquisa de dados e o seu armazenamento - que abrangem diferentes valências de pesquisa nos *corpora*. Para além das motivações que estiveram na origem das FAQ aqui concebidas, a formulação das perguntas evidenciou, simultaneamente, a necessidade de uma intervenção ao nível do *interface* da plataforma TEITOK, complementando, assim, as orientações facultadas pelos materiais disponibilizados.

Para dar conta do desenvolvimento do projeto, este trabalho apresenta-se dividido em duas partes. No primeiro capítulo da primeira parte, esclarece-se o conceito de *corpus* linguístico, destacando-se as vantagens do carácter informatizado e anotado deste tipo de recurso (secção 1.1.). Ainda neste capítulo, tecem-se algumas considerações sobre a aquisição/aprendizagem de língua não materna que servem de moldura ao conceito de *corpora* de aprendentes de L2 e à validade dos dados a extrair destes acervos de dados (secção 1.2.).

A descrição do TEITOK enquanto ferramenta para armazenamento, preparação e pesquisa de dados, tem lugar no capítulo dois. As diferentes potencialidades da plataforma são, aí, descritas em função dos seus utilizadores e, por isso, necessariamente, adotando duas perspetivas distintas, ainda que complementares: a do utilizador interno, aquele que concebe o *corpus* e que procede à sua preparação para futura disponibilização, e a do utilizador externo que recorre à pesquisa para extração de dados.

No final da primeira parte, apresentam-se os *corpora* de aprendentes de português língua não materna que estiveram na génese deste projeto, o PEAPL2_PLE e o COral-Co, salientando a sua afinidade em questões estruturais e metodológicas, mas, também, apontando as suas especificidades em função da natureza distinta das produções que os constituem (cap. 3).

Na segunda parte, descreve-se o trabalho de natureza prática que conduziu à concretização deste projeto. No primeiro capítulo, apresentam-se as principais linhas orientadoras e a metodologia seguida na sua execução, atendendo às várias etapas, desde a definição das FAQ até à sua disponibilização na plataforma *online* TEITOK, bem como os procedimentos observados em cada uma destas etapas.

No início do segundo capítulo, apresentam-se os fundamentos para a definição das FAQ em dois eixos: as áreas de investigação no âmbito do português língua não materna e os diferentes domínios de pesquisa que as FAQ abrangem (secção 2.1.). Seguidamente, apresenta-se um conjunto de FAQ para cada um dos *corpora* (secção 2.2.) e os procedimentos adotados na construção, e disponibilização *online*, das respetivas respostas (secção 2.3.).

Finalmente, com base no trabalho realizado, tecem-se algumas considerações sobre a importância que as FAQ têm, e que poderão ainda vir a ter, na orientação de pesquisa em *corpora* informatizados.

PARTE I - ENQUADRAMENTO

INTRODUÇÃO

Atendendo aos objetivos inerentes ao desenvolvimento do presente projeto, o primeiro capítulo centrar-se-á na definição de *corpus* linguístico, considerando a evolução do termo enquanto simples acervo de dados até ao conceito *moderno* de *corpus* informatizado.

A informatização dos *corpora* foi considerado um marco na história da pesquisa de *corpus*, com especial interesse para a Linguística de *Corpus*, tendo proporcionado inúmeras vantagens, desde a conceção de *corpora* informatizados, passando pelo armazenamento e edição de dados, até à sua pesquisa *online*. Mas se a informatização de *corpora* trouxe consigo a possibilidade de armazenar grandes quantidades de dados e a facilidade e rapidez de pesquisa, não é menos verdade que em muito contribuiu para avanços notáveis na anotação linguística de textos.

O interesse pelos *corpora* informatizados estendeu-se a várias áreas da Linguística, tendo levado à criação de diversos tipos de *corpora*, entre os quais os *corpora* de aprendentes, nos quais incidirá a nossa descrição. Assim, no segundo capítulo, começaremos por tecer breves considerações no âmbito da aquisição/aprendizagem de língua não materna que apontam para questões de investigação nesta área. Seguidamente, descreveremos (secção 2.2.) com algum detalhe os critérios externos e internos que estão subjacentes à criação de *corpora* de aprendentes de língua não materna. O objetivo desta descrição é fundamentar a validade dos dados a extrair de *corpora* de aprendentes e os eventuais contributos da pesquisa de *corpora* informatizados nas áreas da investigação, pedagogia, didática e informática em aquisição/aprendizagem de língua não materna.

Dado que os *corpora* informatizados de aprendentes de português língua não materna que estão na origem deste projeto foram desenvolvidos a partir do TEITOK, esta ferramenta de *software* será alvo de uma descrição detalhada (secção 2.3.), contemplando as suas valências enquanto instrumento para criação de *corpus* e edição de dados, ao serviço do seu utilizador interno, mas também enquanto plataforma de pesquisa *online*, vocacionada, essencialmente, para o utilizador externo. Serão apontados os seus pontos fortes, mas também as suas fragilidades, uma vez que ambos contribuíram para o desenvolvimento do trabalho no âmbito deste projeto.

Por fim (secção 2.4.), proceder-se-á à apresentação dos dois *corpora* de aprendentes sobre os quais incidiu a exploração de valências disponibilizadas pela plataforma TEITOK: o *corpus* de Português Língua Estrangeira (PEAPL2_PLE) e o *Corpus* Oral de Português L2 (COraL-Co), dos quais

destacaremos a estrutura bem documentada e a disponibilização de dados anotados para pesquisa *online*.

CAPÍTULO 1 - *CORPUS* LINGUÍSTICO E *CORPUS* DE APRENDENTES DE L2

INTRODUÇÃO

O capítulo 1 será dedicado, numa primeira secção, à definição de *corpus* linguístico, um conceito que ganha novos contornos na era digital. Com a evolução dos meios informáticos, o acervo de dados alojado de forma rudimentar, em caixas de papel ou gavetas, etiquetados e pesquisados manualmente deu lugar ao *corpus* informatizado. O que tem vindo a mudar de lá para cá?

Se a essência se mantém, no que se refere aos critérios externos e internos que presidem à criação do *corpus*, o que se altera profundamente é a *forma de fazer* e as implicações que esta nova abordagem aporta à pesquisa de *corpora*.

Inicialmente, serão descritos os critérios externos e internos que estão subjacentes à criação de *corpora* linguísticos e, implicitamente, o contributo dos metadados e do sistema de anotação linguística na pesquisa de *corpora* decorrentes da definição destes critérios.

Os critérios externos, tais como a dimensão, a representatividade, a amostragem e o equilíbrio, são essenciais na definição da estrutura do *corpus* e remetem para a importância dos metadados na interpretação dos dados a extrair da pesquisa.

Quanto à anotação linguística, será feita uma breve reflexão em torno dos diferentes níveis de anotação com base nas estruturas da língua, apontando as vantagens deste procedimento para a pesquisa de *corpora*, mas também relembrando os constrangimentos que estão na sua génese.

Já a secção 1.2 será dedicada aos *corpora* de aprendentes de língua não materna: descrição das suas características distintivas, exploração das valências de uma ferramenta digital para criação, edição e pesquisa de *corpus* e apresentação de dois *corpora* de aprendentes de português língua não materna.

Primeiramente, serão tecidas breves considerações sobre o contexto de aquisição/aprendizagem de uma língua não materna, com especial incidência no seu ensino em situação formal de aprendizagem, com o objetivo de dar a conhecer algumas das questões implicadas neste processo e o contributo que os *corpora* de aprendentes podem oferecer nesta área de investigação.

De seguida, serão elencadas algumas das características que definem este tipo de *corpora*, nomeadamente, os critérios subjacentes à sua criação, o perfil do aprendente de língua não materna, a natureza dos dados recolhidos e o seu carácter informatizado.

1.1. A modernidade do *corpus* linguístico: do acervo de dados ao *corpus* anotado

Apesar de a expressão *corpus linguistics* ter surgido nos inícios dos anos 80 do século XX (McEnery *et al*, 2006), já era possível observar uma metodologia na recolha de dados e no seu armazenamento, ainda que em “shoeboxes filled with paper slips” (McEnery *et al*, 2006:3). Talvez por esta razão, os dados assim recolhidos fossem ainda considerados apenas «simple collections of written and transcribed texts» (McEnery *et al*, 2006:3). Parece ser consensual o uso da expressão “*collection of*” para fazer referência aos dados reunidos num *corpus*, numa fase anterior ao interesse da Linguística pela análise de *corpora*, tal como afirmam O’Keeffe & McCarthy (2010): «the term *corpus* had long been in use to refer to a collection or binding together of written works of a similar nature» (O’Keeffe & McCarthy, 2010:5).

Para ilustrar esta época, em que surgiam grandes compilações de dados¹, mas em que não existiam, ainda, os meios para extrair informações de forma consistente, Sardinha (2000) faz alusão aos trabalhos de Thorndike (1921), West (1953) e Quirk (1953)². Este último, o *corpus Survey of English Usage*, viria a servir de modelo ao *corpus* Brown (Brown, 1964), “*the first modern corpus*” (McEnery *et al*, 2006:4). O *corpus* Brown recebeu este epíteto por ter sido um dos primeiros *corpora* informatizados, marcando, assim, um ponto de viragem no que diz respeito à abordagem que é feita de um *corpus*, desde a sua conceção e organização até à sua exploração. Este *corpus* contribuiu, assim, para uma definição mais detalhada e criteriosa, ainda que, por vezes, dada a variedade de *corpora* existentes, essa definição devesse ser “somewhat vague and inclusive term” (McEnery *et al* 2006:5) para não excluir alguns acervos³ resultantes da recolha de dados de outros *corpora*. Desde então, as várias definições de *corpus* linguístico pressupõem o seu carácter informatizado, como é o caso da definição proposta por Sinclair (2004) que a seguir apresentamos:

A corpus is a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research (Sinclair, 2004:16).

¹ Em Portugal, destacam-se como primeiros trabalhos com *corpora*: o projeto Português Fundamental (Lindley Cintra, 1970), o *Corpus* de Frequência (Bacelar do Nascimento, 1978) e o *Corpus* de Referência do Português Contemporâneo (1988).

² O *corpus* SEU (*Survey of English Usage*), compilado por Randolph Quirk só viria a ser disponibilizado em formato informatizado em 1989. Ainda antes, a parte referente aos dados orais foi informatizada dando origem ao *London-Lund Corpus*. Também a *Comprehensive Grammar of the English Language* (1985) de Quirk, Greenbaum, Leech e Svartvik foi baseada no *corpus Survey of English Usage*.

³ Sinclair (2004) dá alguns exemplos daquilo que não deve ser considerado um *corpus*, de acordo com os critérios estabelecidos para a definição de *corpus*. A *World Wide Web*, arquivos, citações, excertos, um texto não constituem *corpora*. No entanto, como argumenta McEnery (2006), o facto de um *corpus* ser especializado ou constituir o *subcorpus* de outro *corpus*, não exclui estes acervos de dados da definição de *corpus*.

Depreende-se das palavras de Sinclair que um *corpus* não se limita a ser uma mera coleção de textos. A sua definição aponta para três aspetos que devem ser tidos em conta quando se fala de *corpus*: (i) o carácter informatizado, (ii) critérios que presidem à criação e (iii) validade dos dados para a pesquisa linguística.

(i) Carácter informatizado

A definição de *corpus* proposta por Sinclair (2004) remete para o seu carácter pesquisável e informatizado⁴. A bem da verdade, o grande incremento na pesquisa de *corpus* deu-se com a possibilidade de informatizar os *acervos*. É indiscutível o contributo das novas tecnologias de que hoje dispomos para a pesquisa de *corpus* e não é, por isso, difícil elencar algumas das vantagens nesta área, como a economia de tempo, a consistência dos resultados obtidos e a possibilidade de *manusear* os dados de acordo com os objetivos de pesquisa (listar, ordenar, armazenar). Estas são, inegavelmente, vantagens que se traduzem, de uma forma generalizada, na facilidade da pesquisa, tornando-a um processo mais rápido e eficiente, permitindo a exequibilidade de respostas a certas perguntas de investigação.

Estas vantagens decorrem do facto de a informatização de *corpora* ter respondido à necessidade de alojar *acervos* de grandes dimensões e de extrair dados viáveis para pesquisa.

Para além das vantagens enunciadas, que, por si só, são inestimáveis, as novas tecnologias permitiram elevar a pesquisa para outro nível ao possibilitarem a anotação linguística dos textos dando origem a *corpora* anotados⁵. A anotação de um *corpus* pode ser feita a dois níveis distintos: a nível estrutural e a nível textual.

No primeiro caso, a anotação compreende os dados externos relativos ao texto, isto é, identificação do *corpus* (autor, ano, dimensão, disponibilização do *corpus*), tipologia textual e descrição das convenções de transcrição e dos níveis de anotação, por exemplo. Estes dados constituem os metadados e são definidos em função da especificidade de cada *corpus* e dele fazem

⁴ Para McEnery, Xiao & Tono (2006:6), «Machine-readability is a de facto attribute of modern corpora».

⁵ Apesar do sistema de anotação de *corpus* já existir, este era um trabalho árduo e moroso do qual dificilmente se retiravam todas as potencialidades por não existirem motores de busca computadorizados que permitissem realizar pesquisas consistentes e rápidas. Aliás, uma das principais críticas apontadas, na época, prendia-se com este facto: «Using paper slips and humans hands and eyes, it was virtually impossible to collate and analyze large bodies of language data» (McEnery *et al*, 2006:4).

parte, quer no cabeçalho dos textos quer em páginas facilmente acessíveis. De uma maneira geral, a inserção de metadados no *software* que aloja o *corpus* permite a sua pesquisa, fornecendo informações importantes na interpretação dos dados. Sobre a importância dos metadados será feita uma reflexão, mais adiante, neste capítulo.

Já a anotação dos textos consiste em atribuir etiquetas de natureza linguística às unidades de significado que constituem o texto, de acordo com categorias previamente definidas nas áreas da morfossintaxe, da semântica, da fonologia ou da pragmática, por exemplo. Atendendo aos objetivos inerentes à conceção de um *corpus*, é possível disponibilizar os textos com diferentes tipos de anotação linguística. Assim, para cada unidade de significado presente no texto pode associar-se uma etiqueta contendo informação em vários domínios da estrutura da língua. Apresentamos, a título de exemplo, alguns dos domínios para os quais é possível criar sistemas de etiquetas e a informação que poderão fornecer⁶:

- a) Morfossintaxe: identificação das classes e subclasses de palavras;
- b) Semântica: desambiguação de significados de palavras polissémicas;
- c) Fonologia: associação da leitura fonética, identificação de elementos paralinguísticos e extralinguísticos;
- d) Pragmática: identificação de conectores discursivos, de modalidades do discurso e de atos ilocutórios.
- e) Lexical: associação de um lema a cada unidade de significado.

Acerca da anotação linguística em *corpora* Leech (2014) defende um maior equilíbrio entre os diferentes níveis de anotação, ainda que flexível, dependendo do tipo de *corpus*, e afirma que a atenção dada aos diferentes níveis de anotação tem sido muito “*patchy*” (Leech, 2014: 25), tal como se pode observar pela listagem, muito elementar, que esboçou do que se tem feito na área da anotação linguística.

⁶ Leech (2004) apresenta uma tipologia de anotação, acrescentando, por exemplo, a anotação estilística e discursiva.

<i>Linguistic level</i>	<i>Annotations carried out so far</i>
phonetic/phonemic	widespread in speech technology corpora or databases
syllabic	none known
morphological	none known
prosodic	little (the LLC and SEC are notable exceptions – see Note 4)
word class (i.e. grammatical tagging)	widespread
syntactic (i.e. parsing)	rapidly becoming more widespread
semantic	none known
pragmatic/discourse	little – but developing

Figura 1: Presença de diferentes níveis de anotação linguística em *corpora* linguísticos (Leech, 2014: 24-25).

Embora a informatização dos *corpora* tenha vindo a facilitar o processo de anotação linguística e, numa primeira fase de preparação dos dados, relativamente a alguns níveis de anotação (por exemplo, a lematização e a classificação morfossintática), seja possível fazê-lo automaticamente, este processo não dispensa um trabalho manual *a posteriori* de verificação e retificação, sempre que necessário, de falhas de etiquetagem. Saliente-se, no entanto, que, nos casos da anotação de natureza semântica, fonológica e pragmática, este processo terá que ser realizado previamente de forma manual, uma vez que os *software* disponíveis não permitem, ainda, desambiguar contextos, reconhecer atos discursivos ou proceder à transcrição fonética. A verificação manual, embora morosa, é essencial, pois só desta forma é possível alcançar consistência nos resultados.

O processo de anotação deve preservar sempre a integridade do texto original e a anotação deve poder ser consultada paralelamente a este, mas nunca se sobrepor ou substituí-lo. Não podemos esquecer que o processo de anotação não está isento de parcialidade, pois tal como refere Leech (2004), «Corpus annotation is the practice of adding interpretative linguistic information to a corpus»⁷ (Leech, 2004:17). As opções de classificação refletem quadros teóricos de quem leva a cabo este processo e, por isso, é tão importante dar a conhecer ao utilizador do *corpus* as operações efetuadas no texto a este nível.

⁷ Talvez por esta razão, existam *corpora* que, apesar de informatizados, não são anotados.

Ainda assim, também não é menos verdade que a anotação dos textos é um instrumento inestimável na pesquisa de *corpora* porque permite obter resultados de pesquisa que se estendem muito para além das frequências de ocorrências em diferentes contextos. A anotação possibilita a restrição da pesquisa em função de hipóteses formuladas pelo utilizador ao nível dos diferentes níveis de estruturação da língua, mas também, e sobretudo, em função da reformulação dessas mesmas hipóteses na sequência dos resultados obtidos, numa espécie de interação entre o *corpus* e o investigador: «From this perspective, probably a majority view, adding annotation to a corpus is giving 'added value'» (Leech, 2004:17).

Podemos afirmar, então, que a grande vantagem da informatização dos *corpora* está não só na forma como possibilita armazenar grandes quantidades de dados, mas essencialmente na forma como permite organizá-los para subseqüentemente os disponibilizar para pesquisa.

(ii) Critérios que presidem à criação

As informações que constituem a documentação do *corpus*, e sem as quais não seria possível interpretar fielmente os dados obtidos, constituem os metadados, ou seja, 'data about data' (Burnard, 2004). Sem estes, o investigador apenas teria acesso a um conjunto de dados desconexos e descontextualizados. É também esta a ideia reforçada por Freitas (2015):

A documentação possibilita (i) avaliação relativa à adequação do material às questões de pesquisa e a conseqüente reutilização do material, e (ii) interpretação consistente dos resultados (Freitas, 2015: 34).

Os metadados constituem a moldura do *corpus* e deverão constar dele, em secção apropriada para o efeito, e serem facilmente acessíveis, mediante pesquisa ou remissões. Estes podem organizar-se em dois níveis distintos, de acordo com os critérios que presidem à criação do *corpus*: os critérios externos e os critérios internos.

É preciso, desde logo, clarificar para que serve o *corpus*, qual é o objetivo que está na sua génese, que tipo de textos integrará (textos orais/escritos, pertencentes a determinado(s) género(s) textual(-is), correspondendo ao registo formal/informal), qual a fonte e o ano a que se referem os textos, como serão disponibilizados os dados. As respostas a estas questões permitirão balizar o trabalho a realizar *a posteriori* e pré-determinar, por força do tipo de *corpus* que se pretende construir, os critérios externos.

Como já se referiu, a *dimensão*, a *representatividade*, a *amostragem* ou o *equilíbrio* do *corpus* constituem critérios externos (Sinclair, 2004) que se podem estabelecer aquando da conceção de um *corpus*. Estes conceitos têm sido amplamente discutidos na literatura (Sardinha, 2000; Sinclair, 2004; McEnery, Xiao & Tono, 2006) e ainda que não tenha lugar, no âmbito deste trabalho, uma reflexão aprofundada sobre a forma como estes critérios se refletem na criação de *corpora*, impõem-se algumas observações de natureza mais abrangente.

A dimensão e a representatividade do *corpus* são dois critérios de difícil distinção, uma vez que estão intimamente relacionados. Como a representatividade de um *corpus* é muito subjetiva e, por isso, difícil de definir, muitas vezes, este critério associa-se à dimensão, pensando-se que a representatividade de um *corpus* é tanto maior quanto maior for a sua dimensão. No entanto, como alerta Sardinha, «A representatividade está ligada à questão da probabilidade» (Sardinha, 2000:343), ou seja, um *corpus* pode ser extenso e não ser representativo de uma determinada língua ou variedade, da mesma forma que um *corpus* menos extenso pode satisfazer este critério. A reflexão de Fillmore (1992) sobre a importância dos *corpora* na análise linguística reflete a natural indefinição que envolve os conceitos de representatividade e de dimensão:

[...] I don't think there can be any corpora, however large, that contain information about all of the areas of English lexicon and grammar that I want to explore; [...] every corpus that I've had a chance to examine, however small, has taught me facts that I couldn't imagine finding out about in any other way (Fillmore, 1992:35).

Tal acontece porque é difícil prever a ocorrência de determinados dados no *corpus*, mas também porque «A corpus that sets out to represent a language or a variety of a language cannot predict what queries will be made of it» (Sinclair, 2004:6), como ressalva Sinclair. Por esta razão, o autor defende que a única forma de sustentar a representatividade de um *corpus* é alicerçar a sua conceção através da definição inequívoca dos objetivos que estão na sua génese.

Quanto à amostragem e ao equilíbrio do *corpus*, estes critérios prendem-se com a tipologia e a quantidade de textos de cada tipo a inserir. É preciso salientar que o tipo e a quantidade de textos que integram um *corpus* devem reger-se, principalmente, por critérios de natureza externa, isto é, um acervo deve ser concebido em função da amostra que se pretende representar e não em função dos dados linguísticos que se pretendem encontrar. Relembramos que um *corpus* é apenas uma amostra, um recorte, e como tal não deve pretender ser exaustivo.

Se numa fase inicial de conceção do *corpus*, os critérios externos são essenciais na definição da sua estrutura, posteriormente, é necessário estabelecer os critérios internos. Os critérios internos, por sua vez, estão diretamente relacionados com aspetos de natureza linguística respeitantes aos textos recolhidos.

Assim, ao conceber um *corpus*, é preciso definir os procedimentos a ter em conta na preparação dos textos que o constituem. Dependendo da especificidade de cada *corpus*, as tarefas, neste âmbito, poderão ser variadas. Entre elas, destacamos, por exemplo, armazenar e identificar os textos (por autor/fonte, ano, área/domínio/tarefa), “limpar” os textos⁸, digitalizar ou transcrever textos, preparar ficheiros áudio, no caso de *corpora* orais, ou anotar os textos. Este trabalho é bastante meticuloso e para o levar a cabo é necessário estabelecer aquilo a que poderíamos chamar *subcritérios*. Na verdade, estes *subcritérios* resultam do apuramento de procedimentos que deverão ser observados na realização das tarefas relacionadas com a preparação dos textos: Como identificar os textos? O que remover dos textos? Que convenções de transcrição utilizar? Que tipo de anotação realizar? Como segmentar os ficheiros áudio? A definição de *subcritérios* deve ser rigorosa porque deles depende a consistência da informação a extrair do *corpus*.

Todas estas informações relativas à conceção do *corpus*, e que resultam da ponderação de critérios externos e internos, devem constar do *corpus* e ser do conhecimento do utilizador. A forma como esta informação é disponibilizada é variável, podendo acompanhar os textos, na forma de cabeçalho, ser apresentada em diferentes secções ou páginas do *corpus* ou em documentos armazenados para consulta.

Um dos grandes avanços resultantes da informatização também se verifica neste domínio, pois esta veio permitir a pesquisa não só dos dados, mas também dos metadados, fortalecendo a relação existente entre os critérios externos que estruturam os *corpora* e os critérios internos que permitem extrair informação de natureza linguística, tornando mais explícita a importância que os metadados, informação resultante da definição dos critérios externos e internos, têm na interpretação dos dados.

(iii) Validade dos dados para a pesquisa linguística

O conceito de *corpus* que apresentámos «as a source of data for linguistic research» (Sinclair, 2004) está, como já vimos, diretamente relacionado com a disponibilização de *corpora*

⁸ Por vezes, é necessário fazer um trabalho de limpeza, removendo determinados tipos de formatação que impedem a correta leitura pelos programas informáticos utilizados para o processamento dos textos, ou informação irrelevante para o *corpus* linguístico (por exemplo, imagens, gráficos).

informatizados. E foi esta nova concepção que veio despertar o interesse da Linguística de *Corpus* (Sardinha, 2000), que vê definido o seu objeto de estudo, nas palavras de Mendes, da seguinte forma:

A Linguística de *Corpus* baseia o estudo da língua em ocorrências extraídas de um *corpus*, isto é, de um conjunto de textos escritos (ou excertos de textos) ou de transcrições de registos orais, tipicamente em formato electrónico (Mendes, 2016:224).

Mas o que torna válidos os dados extraídos de *corpora* informatizados para o estudo de uma língua ou variedade de uma língua? Em nosso entender, a objetividade e a consistência dos dados que podem ser extraídos de *corpora* linguísticos, atendendo às características já apresentadas em (i) e (ii), nomeadamente o potencial dos meios informáticos para armazenamento e extração/pesquisa de informação e os critérios subjacentes à criação de um *corpus*, são duas características essenciais para a validade dos dados para pesquisa linguística.

Independentemente da perspectiva do investigador e da teoria linguística que a suporta⁹, a pesquisa num *corpus* informatizado possibilita a extração de dados de forma objetiva e consistente. Tal acontece graças a programas informáticos que permitem, no contexto da pesquisa *online*, contar palavras, listar ocorrências em contexto e dar conta da frequência com que ocorrem os dados. Este tipo de pesquisa permite observar, por exemplo, a sistematicidade da língua, tal como Simone Sarmiento (2010) a descreve:

A variação sistemática, ou seja, a recorrência de traços linguísticos (colocação, coligação, padrão sintático, entre outros) indica que a linguagem é padronizada (*patterned*) e motivada por diversos fatores além das necessidades comunicativas (Sarmiento, 2010:89).

O investigador tem, assim, a possibilidade de observar, registar e interpretar os dados empíricos de forma objetiva, dados esses que ilustram diversos fenómenos que ocorrem na língua e que não seriam observáveis apenas com base no conhecimento intuitivo que os falantes nativos têm da sua língua.

⁹ O *corpus* informatizado, e toda a estrutura que o envolve, tornou-se uma ferramenta indispensável ao estabelecer uma interação com o utilizador que lhe permite confirmar as suas hipóteses através dos dados (*corpus-based*), mas também levantar questões, inicialmente insuspeitas, em função dos dados obtidos (*corpus-driven*).

Aliada à objetividade, os *corpora* linguísticos reúnem textos autênticos, isto é, textos que não foram artificialmente produzidos para integrar o *corpus*. Assim, estes fornecem informações sobre fenómenos diversos que ocorrem na língua, tornando-os um instrumento válido para a descrição e análise de dados empíricos, favorecendo «a discussão de questões teóricas solidamente fundamentadas» (Bacelar, 2002) e, conseqüentemente, um melhor conhecimento das línguas.

A grande vantagem da pesquisa *online* de *corpora* linguísticos é disponibilizar os dados à medida dos seus utilizadores, independentemente do fim a que se destinam. Queremos com isto dizer que a pesquisa não é estanque, não é fechada sobre si mesma ou centrada num único objetivo, isto é, os dados continuam a ser válidos para análise linguística em diferentes áreas do conhecimento, dependendo de «the research question being investigated» (McEnery, Xiao & Tono, 2006:121).

O que importa aqui referir não é a multiplicidade de *corpora* existentes, mas antes o que torna os dados de *corpora* linguísticos válidos para a pesquisa e um contributo inegável em áreas da linguística tão diversas como o léxico, a gramática, a semântica, a análise do discurso, a tipologia textual, os estudos literários, a tradução, a investigação, a pragmática, a sociolinguística, as variedades dialetais, os estudos contrastivos e o ensino/aprendizagem de línguas (McEnery, Xiao & Tono, 2006; O'Keeffe & McCarthy, 2010). Poderemos acrescentar, ainda, a esta extensa lista, a criação de materiais didáticos e a criação e atualização de instrumentos tecnológicos. Por outras palavras, segundo O'Keeffe & McCarthy (2010:9), «CL has had much to offer other areas by providing a better *means* of doing things».

Uma das áreas em que a análise de *corpus* ganhou especial relevância foi a da aquisição/aprendizagem de língua não materna (LNM). Neste sentido, foram criados *corpora* de aprendentes com o principal intuito de apoiar a descrição da forma como o aprendente adquire/aprende uma língua não materna.

Se, nesta secção, descrevemos, com algum detalhe, o conceito de *corpus*, no sentido mais lato do termo, cabe agora restringi-lo em função dos *corpora* de aprendentes de L2 que constituem o objeto de trabalho deste projeto e que serão apresentados no capítulo 3.

1.2. *Corpora* de aprendentes de L2

1.2.1. Considerações prévias: aquisição e aprendizagem de L2

No intuito de abordar o conceito de *corpora* de aprendentes de L2 e a sua importância para a Linguística de *Corpus*, é necessário tecer algumas considerações genéricas, mas relevantes no contexto de aquisição/aprendizagem de uma língua não materna.

Antes de mais, é necessário esclarecer que o presente projeto foi desenvolvido a partir de *corpora* de língua não materna, que serão descritos mais adiante (cap. 3), e que, atendendo ao perfil dos aprendentes, cujas produções fazem parte destes acervos, e ao contexto de aprendizagem da língua não materna, utilizaremos, doravante, a expressão L2 como equivalente de língua não materna¹⁰.

Sabemos, à partida, que o perfil sociolinguístico dos aprendentes é muito heterogêneo (idade, sexo, proficiência linguística em línguas estrangeiras, países em que já viveu) e que pode condicionar as suas aprendizagens no domínio da língua não materna. A idade pode ser, desde logo, um fator preponderante na aprendizagem da língua estrangeira, especialmente se tal acontece já na adolescência ou na fase adulta, no chamado *período crítico* (Johnson & Newport, 1989). Nestes casos, o aprendente *tardio* apresentará dificuldades na aprendizagem da língua não materna, sobretudo na área da fonética/fonologia, denunciadas pelo “*sotaque*” estrangeiro. Ativará, no entanto, outros mecanismos de aprendizagem, recorrendo, essencialmente à sua memória declarativa (Martins, 2008).

Não menos importante é a sua motivação e a sua aptidão para a aprendizagem de uma língua estrangeira, decorrentes de fatores externos (motivação profissional, contexto de aprendizagem, país de origem), bem como a sua própria personalidade (o aprendente é, ou não, organizado, introvertido, aplicado).

Quando inicia o processo de aprendizagem, o aprendente quer rapidamente começar a *falar* a língua que está a aprender. Desde cedo, é possível observar quais as estruturas que adquire primeiro e começa a utilizar com maior ou menor dificuldade. Mas também é neste momento que a língua

¹⁰ O conceito de LE (Língua Estrangeira) é frequentemente utilizado para fazer referência a aprendentes que aprendem uma língua não materna em contexto de instrução formal, e que, cumulativamente, não têm qualquer contacto com esta língua fora da sala de aula (Flores, 2014) No entanto, os *corpora* que mais à frente apresentaremos incluem produções de aprendentes que, embora se encontrem a aprender a língua em contexto formal de aprendizagem, também se encontram no país em contexto de imersão.

materna (L1) pode interferir no processo de aprendizagem da L2, através de transferências negativas ou positivas de vocabulário ou de estruturas morfosintáticas e fonológicas da L1 para a L2. É normal que o aprendente tenha momentos de hesitação perante estruturas mais complexas, ou ambíguas, da língua-alvo (LA) e use estratégias de evitamento para as ultrapassar. Nesta fase, é preciso observar os desvios e valorizar a sua ocorrência na tentativa de perceber e interpretar o modo como o processo se está a desenrolar. À medida que ganha alguma segurança, o aprendente já é capaz de se autocorrigir e reformular os seus enunciados.

Todos estes aspetos - a idade, a motivação, o conhecimento prévio, os padrões de aquisição, a transferência positiva/negativa, as estratégias de evitamento, a autocorreção - refletem-se no uso que o aprendente faz da língua - o *output*. Note-se que a forma como utiliza a língua, seja em contexto formal de aprendizagem, seja, por exemplo, em contexto de imersão (ou na combinação dos dois), revela os vários estádios no desenvolvimento de competências na L2. Refira-se, ainda, que estes estádios podem não ser (e usualmente não o são) coincidentes nos diferentes domínios da língua, isto é, o aprendente pode encontrar-se num determinado nível de proficiência no que se refere ao domínio da comunicação escrita, mas revelar outro nível de proficiência ao nível da comunicação oral, por exemplo. O aprendente constrói, assim, ao longo da sua aprendizagem, o(s) seu(s) próprio(s) sistema(s) linguístico(s), a sua interlíngua, com características muito próprias (Selinker, 1972; 2014).

Nem sempre o processo de aquisição/aprendizagem de uma LNM termina quando o aprendente utiliza com segurança as estruturas da LA e faz delas uso adequado em contexto. Por vezes, o aprendente continua a manifestar dificuldades na pronúncia de alguns segmentos fonológicos ou na correta realização de determinadas estruturas da língua. Será que tal acontece devido a fatores inerentes ao próprio aprendente (desistiu, não investe porque o nível em que se encontra já satisfaz os seus objetivos, tem dificuldades de aprendizagem, é aprendente *tardio*) ou a fatores inerentes à própria língua-alvo (como a ambiguidade, por exemplo)? Quando tal acontece, ocorre a fossilização.

Face ao que sumariamente se expôs, podemos afirmar que as características do processo de aquisição/aprendizagem de uma língua não materna são observáveis e a sua descrição sustentada empiricamente, com recurso a dados de *corpora* de aprendentes de L2.

1.2.2. Conceito

Um *corpus* de aprendentes, embora partilhe algumas características com os demais *corpora* linguísticos, já retratados na secção 1.1., no que se refere aos critérios que presidem à sua criação, recolha, preparação e disponibilização de textos, «requires clarification both as regards the status of the speakers involved and the type of data they produce» (Granger, 2008:1).

Granger, Gilquin & Meunier (2005:9) definem *corpora* de aprendentes como «eletronic collections of natural or near-natural data produced by foreign or second language (L2) learners and assembled according to explicit design criteria». Esta definição contempla, para além (i) do carácter informatizado do *corpus*, três outros aspetos distintivos: (ii) o tipo de dados recolhidos, (iii) dados produzidos por aprendentes de L2 e (iv) critérios específicos inerentes à conceção do *corpus*. Por uma questão metodológica, abordaremos estes aspetos pela ordem inversa: (i) critérios específicos inerentes à conceção do *corpus*; (ii) dados produzidos por aprendentes de L2; (iii) o tipo de dados recolhidos e (iv) carácter informatizado do *corpus*.

(i) Critérios específicos inerentes à conceção do *corpus*

No capítulo intitulado *From design to collection of learner corpora* (Granger, Gilquin & Meunier eds, 2005:9-34), Gaëtanelle Gilquin reúne alguns critérios que considera serem distintivos na criação de um *corpus* de aprendentes, contribuindo para definir a sua estrutura e aferir os critérios externos que o sustentam - nomeadamente, a amostra, a representatividade, a dimensão e o equilíbrio do *corpus*, como havíamos já referido anteriormente. São propostos, então, os seguintes critérios:

A. Tipologia do corpus

Quanto à tipologia, um *corpus* pode ser oral, escrito (também se inclui nesta categoria o *corpus* constituído apenas por transcrições de produções orais) ou misto.

B. Tipologia textual

Relacionada com a amostra e com o equilíbrio do *corpus*, a tipologia textual deve ser definida criteriosamente, seleccionando-se os tipos de textos a incluir no *corpus* escrito (carta formal / informal, textos de opinião, textos argumentativos, descritivos, relatórios) ou no *corpus* oral (entrevista, leitura, diálogos) e o número de textos de cada tipo.

C. Língua materna e língua-alvo

O *corpus* pode ser constituído por produções de aprendentes que partilham a mesma língua materna (“mono-L1”) ou de aprendentes com diferentes línguas maternas (“multi-L1”) que aprendem, geralmente, uma língua-alvo, podendo também aprender várias línguas estrangeiras (multi-L1/multi-L2), embora menos frequente.

D. Recorte temporal

No que diz respeito aos *corpora* de aprendentes de L2, este é um critério sobre o qual importa tecer algumas considerações. É possível estabelecer um período de tempo que balizará a recolha de produções, dando origem a um *corpus* de tipo sincrónico. Neste caso, os dados são recolhidos em função dos diferentes níveis de proficiência, num determinado momento temporal, constituindo um recorte transversal.

Outra opção é recolher produções ao longo de distintos períodos de tempo, acompanhando o aprendente ao longo do processo de aquisição/aprendizagem, criando assim um *corpus* diacrónico. Um *corpus* de tipo diacrónico é considerado um *corpus* longitudinal, que permite observar as várias fases de desenvolvimento de competências na L2, ao longo do tempo, por parte do aprendente. No entanto, a criação deste tipo de *corpus* depara-se com importantes obstáculos, desde logo pela dificuldade em acompanhar o percurso de um aprendente que, por vezes, desiste a meio do percurso ou não lhe dá continuidade, interrompendo o processo de recolha de dados.

Perante esta dificuldade, há a possibilidade de criar «quasi-longitudinal corpora¹¹ (i.e. corpora gathered at a single point in time but from learners of different proficiency levels)» (Granger, 2004:131) que, não sendo *corpora* longitudinais, na verdadeira aceção do termo, podem, ainda assim, contribuir com dados relevantes para o estudo do desenvolvimento das interlínguas.

E. Raio de abrangência

Os *corpora* de aprendentes de L2 podem ter um raio de abrangência territorial/espacial mais vasto, a nível nacional, ou limitar-se a uma região ou um local do país. Na verdade, no que concerne aos *corpora* de aprendentes, a tendência é para que sejam *locais*, visto que, por vezes, a criação do *corpus*, e a posterior recolha de dados, é realizada por professores / investigadores com o objetivo de identificar «one’s own learners’ specific needs through a corpus analysis of their output and thus provide tailor-made solutions to their problems» (Gilquin, 2005: 5-6).

F. Finalidade e Disponibilização

¹¹ Gilquin, parafraseando Gass e Selinker (2008:56–7), refere-se à utilização da expressão “corpora of pseudolongitudinal data” pelos autores para designar este tipo de *corpus*.

Mediante a(s) finalidade(s) do *corpus*, também a sua disponibilização pode diferir. Nos casos em que são concebidos por entidades comerciais, com vista à posterior elaboração de materiais, os *corpora* não estão, muitas vezes, acessíveis ao público em geral. Contrariamente a estes, aqueles que são criados com objetivos académicos, relacionados com a investigação «in educational settings and interested in learning more about interlanguage (possibly with pedagogical aims in mind)» (Gilquin, 2005:6) e cuja divulgação se faz através da partilha com a comunidade científica são, regra geral, de livre acesso ao público.

Para além destes critérios externos, específicos da criação de um *corpus* de aprendentes, julgamos que uma outra decisão deve ser tomada nesta fase: a decisão de anotar ou não o *corpus*. É importante decidir se o *corpus* será ou não anotado, mesmo que esta etapa não seja realizada de imediato, porque esta decisão pode condicionar a forma como os dados são armazenados e preparados para serem disponibilizados.

Relembramos que toda a informação referente aos critérios externos que estiveram na origem do *corpus* deverá estar acessível ao público, dada a importância dos metadados na pesquisa de *corpus* (cf. secção 1.2.1., capítulo1).

(ii) Dados produzidos por aprendentes de L2

O que distingue os *corpora* de aprendentes de outros tipos de *corpora* é, precisamente, o facto de as produções que os integram serem produzidas por aprendentes e, inevitavelmente, serem condicionadas por fatores como o seu conhecimento linguístico prévio e o nível de proficiência. Neste contexto, ganham especial relevância os metadados referentes ao perfil do aprendente.

Com base no trabalho de Ellis (1994), Granger (2008:4-5) apresenta algumas das variáveis que podem influenciar a aprendizagem de uma LNM e que, por isso, devem ser tidas em conta na caracterização do perfil do aprendente. Algumas variáveis como a idade, o sexo, a nacionalidade ou a língua materna podem ser consideradas informações de carácter mais genérico, mas a estas podem acrescentar-se variáveis mais específicas no que respeita a aprendizagem da L2, como o contexto de aprendizagem, o nível de proficiência, o conhecimento linguístico que já possui da L2 ou de outras línguas estrangeiras (cf. atrás, secção 1.2.1.).

Relativamente a este último grupo de variáveis, destacamos a importância, mas também a dificuldade, de recolher informação objetiva acerca do contexto de aprendizagem e do nível de proficiência do aprendente. As informações sobre o contexto de aprendizagem devem esclarecer aspetos como «the type and amount of *input* received in class or during extracurricular activities, the

time spent in a country where the target language is the official language» (Meunier, 2016:3), aspetos estes que são preponderantes no desenvolvimento de competências na aprendizagem de uma L2. Já o nível de proficiência, embora seja regulado por documentos oficiais, pode não corresponder exatamente àquele em que se enquadra o aprendente (condicionantes externas) e pode variar em função dos diferentes domínios da língua, isto é, o nível de proficiência em que o aprendente se enquadra pode resultar de diferentes desempenhos em diferentes domínios da língua (compreensão oral/escrita, produção oral/escrita).

Assim, com vista à recolha de informações, elabora-se, habitualmente, um questionário contemplando questões relacionadas com este tipo de variáveis¹², no qual se inclui uma declaração de consentimento informado, a preencher pelos aprendentes.

Após a compilação dos dados, estes são inseridos numa base com vista à sua disponibilização e com a vantagem, no caso dos *corpora* informatizados, de serem pesquisáveis por si só ou combinados com outros critérios de pesquisa para restringir resultados. Num *corpus* de aprendentes, os metadados respeitantes aos informantes podem acompanhar as produções, no cabeçalho, ou serem disponibilizados em secção destinada para o efeito, facilmente acessíveis através de *links*.

(iii) O tipo de dados (quase-) autênticos

Quanto às produções recolhidas, estas resultam da apresentação de estímulos variados, seja no domínio da escrita ou da oralidade, almejando tanto quanto possível a autenticidade, mas nunca esquecendo que, ao serem produzidas, maioritariamente, em contexto formal de aprendizagem, não será possível aferir se, perante uma situação idêntica em contexto informal, o enunciado produzido seria o mesmo. Ainda assim, o aprendente produz enunciados de acordo com aquilo que considera ser o adequado em determinadas situações comunicativas e esse é o ponto de partida para a extração de dados de um *corpus* de aprendentes.

Em contextos de aprendizagem formal, os dados recolhidos dependem das tarefas que são apresentadas aos aprendentes e das condições proporcionadas para a realização dessas mesmas tarefas.

Relativamente às tarefas com vista à recolha de produções escritas, que integram *corpora* escritos, estas podem ir desde a escrita de textos de apresentação ou de opinião, textos

¹² Além destas, podem ser consideradas outras variáveis, em função do tipo de *corpus* que se pretende criar, tal como Granger escreve, citando Ellis (1994:49): «the factors that can bring about variation in learner output are numerous, perhaps infinite» (Granger, 2008:4).

argumentativos ou descritivos, recontos, relatórios até à redação de cartas formais ou informais, postais ou recados. Já no que se refere aos *corpora* orais¹³, as produções recolhidas resultam, em muitos casos, da eliciação de tarefas. Constituem alguns exemplos o reconto de uma história, lida anteriormente ou a partir de imagens, interação entre pares simulando situações do quotidiano, leitura de texto ou de lista de palavras, entre outras.

Há alguns fatores que se prendem com a realização das tarefas e que, podendo ser controlados pelo investigador, muitas vezes, condicionam o desempenho do aprendente. Eis alguns exemplos: o tempo disponibilizado para a realização da tarefa, o assunto sobre o qual recai a atividade, a possibilidade de consulta de materiais, como dicionários ou gramáticas (em formato impresso ou digital), o facto de a tarefa ser realizada em situação de avaliação ou não, a existência de um tempo previamente definido para a preparação da tarefa a executar, o desempenho de colegas (no caso de tarefas de interação comunicativa) ou, mesmo, constrangimento provocado pelas condições de gravação, quando há lugar à gravação de produções orais, por exemplo.

A descrição sumária das tarefas levadas a cabo pelos aprendentes deve integrar o *corpus* e deve, à semelhança dos dados referentes aos aprendentes, ser pesquisável. É desejável que as produções resultantes da realização de tarefas específicas sejam pesquisáveis através de um campo de pesquisa autónomo, de modo a permitir a consulta de produções em função deste critério e/ou da conjugação de critérios.

(iv) Caráter informatizado do *corpus*

Na secção 1.1., aquando da definição de *corpus* linguístico, observámos que a informatização é uma característica dos *corpora* da era *moderna*. Logicamente, o *corpus* de aprendentes também tira partido deste avanço tecnológico, justificando, aliás, o uso da expressão *Computer Learner Corpora* (Granger, 1998). Granger define *corpora* de aprendentes da seguinte forma: «Computer learner corpora are electronic collections of spoken or written texts produced by foreign or second language learners» (Granger, 2002:124).

Efetivamente, Granger define o *corpus* de aprendentes informatizado como um acervo de dados digital com todas as suas implicações, isto é, as ferramentas digitais estão ao serviço do *corpus* desde a conceção à disponibilização, conferindo rapidez e precisão aos procedimentos, nomeadamente, na

¹³ Ou escritos, nos casos em que há apenas lugar à transcrição e não há a possibilidade de aceder a ficheiros áudio.

recolha de dados, no processo de transcrição, no armazenamento, no desenvolvimento de sistemas de anotação e de motores de pesquisa *online*.

Uma vez que mais à frente, neste trabalho, será apresentada uma ferramenta digital para armazenamento, tratamento e disponibilização de dados de *corpora* de aprendentes - o TEITOK - e serão apresentados dois *corpora* de aprendentes de L2, mais concretamente, de português L2 - PEAPL2_PLE e COral-Co -, a descrição das diferentes etapas, da recolha à pesquisa de dados, será realizada com detalhe, oportunamente, nos respetivos capítulos.

O que queremos aqui salientar é que as vantagens das novas tecnologias destinadas à criação de *corpora* e pesquisa de dados, como anteriormente se antecipou, têm o seu reflexo na investigação: «By offering more accurate descriptions of learner language than have ever been available before, computer learner corpora will help researchers to get more of the facts right» (Granger, 1998:17).

1.2.3. Virtualidades dos *corpora* de aprendentes de L2

Se, de uma maneira geral, a análise de *corpora* favorece a descrição de estruturas linguísticas ao criar a possibilidade de extrair ocorrências referentes a uma língua, ou a variedades de uma língua, a análise de *corpora* (escritos e orais) de aprendentes de língua não materna possibilita «sustentar empiricamente a investigação em torno das características da interlíngua, tal como ela se apresenta em determinados e distintos níveis de proficiência» (Santos *et al*, 2016). No mesmo sentido, Granger (2008) argumenta que os *corpora* de aprendentes são úteis na investigação de questões relacionadas com «the exact role of transfer in second language acquisition and the notion of avoidance» (Granger, 2008:8).

Um *corpus* de aprendentes que englobe sujeitos de diferentes línguas maternas (multi-L1) permitirá, por um lado, descrever o papel destas línguas no processo de aquisição/aprendizagem da L2, ao possibilitar a realização de estudos comparativos sobre a forma como este processo se desenrola em aprendentes com diferentes L1. Por outro lado, permite averiguar quais as áreas consideradas problemáticas na língua-alvo, em função dos resultados obtidos nestes estudos.

Trabalhos de investigação em *corpora* longitudinais, por exemplo, permitirão observar os efeitos de transferência da L1, ou de outras L2, no desenvolvimento de competências na língua não materna, e averiguar qual a influência do *input* em função do contexto de aprendizagem. É também possível descrever as estruturas da L2 que demoram mais tempo a adquirir pelo aprendente, e que estratégias este usa para ultrapassar essas dificuldades, quais as estruturas que são mais facilmente aprendidas em contexto instrucional de aprendizagem ou aquelas que dificilmente chegam a ser aprendidas.

Contudo, a análise de *corpora* de aprendentes não se esgota na descrição sustentada empiricamente do processo de aquisição/aprendizagem de uma L2. Os dados resultantes da análise de *corpora* poderão ser úteis no processo de ensino-aprendizagem, não só fornecendo pistas sobre o que deve ser ensinado em sala de aula, mas também permitindo aos aprendentes observar e discutir os dados, enriquecendo a sua aprendizagem, sempre que estejam reunidas condições para que tal aconteça (nível de proficiência do aprendente, competências de exploração de *corpus* por parte do professor, acesso a computadores na sala de aula, adequação a conteúdos programáticos).

Este tipo de *corpora* também fornece dados válidos para a construção de materiais didático-pedagógicos, como é o caso de dicionários, gramáticas ou manuais para um ensino baseado em tarefas.

Para além destas aplicações, a análise de *corpora* informatizados também acaba por permitir recolher informações para a otimização e rentabilização dos meios tecnológicos, tal como enumera Mendes (2016):

Os corpora, [...] são ainda fonte de informação para a criação de aplicações várias, como, por exemplo, redes conceptuais, sistemas de sumarização automática, de extração de informação, de tradução automática, de reconhecimento da fala e síntese de voz. A área das Humanidades em geral pode beneficiar das metodologias e aplicações desenvolvidas para os *corpora* de língua pela Linguística de Corpus e Linguística Computacional (Mendes, 2016:226).

CAPÍTULO 2 - TEITOK: uma ferramenta para armazenamento, preparação e pesquisa de dados

INTRODUÇÃO

Proceder-se-á, neste capítulo, à descrição das valências de um *software* que possibilita a criação, edição e pesquisa de *corpus*: o TEITOK. O TEITOK é uma ferramenta informática que permite trabalhar com *corpora* desde a sua conceção à sua pesquisa e que apresenta inúmeras vantagens, mas também algumas fragilidades. A descrição pormenorizada desta ferramenta, na perspetiva dos seus utilizadores (interno e externo), prende-se com um dos objetivos do presente projeto: dar a conhecer as principais valências da plataforma TEITOK na exploração de dois *corpora* de PL2: o PEAPL2_PLE e o COral-Co, o que se fará no capítulo 3.

O TEITOK é uma plataforma desenvolvida para trabalhar com *corpora* linguísticos de natureza diversa¹⁴. No âmbito deste projeto, porém, a descrição que aqui faremos incidirá sobre a aplicabilidade do TEITOK, naquilo que são as suas virtualidades, mas também as suas fragilidades, enquanto “*tool design*” (Janssen, 2019), destinada a *corpora* de aprendentes de L2. Até ao momento, o sistema TEITOK permitiu desenvolver os seguintes projetos de *corpora* de aprendentes de L2:

- em Portugal, COPLE2 (Mendes *et al*, 2016), PEAPL2 (Cristina Martins, 2008) e os *subcorpora* PLE, Timor e Guiné-Bissau (ainda em desenvolvimento), COral-Co (Isabel Santos, 2012);
- na Croácia, CrolTeC (Preradovic *et al*, 2015);
- na Lituânia, ESAM (Znotina, 2015?).

No próximo capítulo, apresentar-se-ão dois *corpora* de aprendentes que foram desenvolvidos a partir do sistema TEITOK e que estão na génese do projeto que será descrito na segunda parte deste trabalho: o *Corpus de Produções Escritas de Aprendentes de PL2: Subcorpus Português Língua Estrangeira* (PEAPL2_PLE) e o *Corpus Oral de Português L2 - Coimbra* (COral-Co). Nesse sentido, julgamos ser mais oportuno apresentar algumas das funcionalidades da plataforma, que agora descrevemos, *a posteriori*, partindo de exemplos contextualizados.

Embora não tenha sido concebida, à partida, como uma ferramenta direcionada para trabalhar com *corpora* de aprendentes, o TEITOK possui características que o tornam muito atrativo neste domínio. Isso deve-se ao facto de o TEITOK ser «a web-based framework for corpus creation, annotation, and distribution, that combines textual and linguistic annotation within a single TEI based XML document» (Janssen, 2016:4037). Depreendemos, das palavras de Janssen, que as principais valências da plataforma são (i) a criação de *corpus* informatizado, (ii) a preparação e edição de dados e (iii) a pesquisa *online*, aspetos que exploraremos de seguida.

Quanto a nós, o foco destas valências incide no trabalho a desenvolver com *corpora* em duas perspetivas distintas, ainda que complementares. Se, por um lado, a criação de *corpus* informatizado e a preparação e edição de dados é um trabalho que cabe ao criador do *corpus*, ao seu utilizador interno, portanto, já a pesquisa, embora desenhada por este, e concebida pelo administrador da plataforma, tem como principal alvo o utilizador externo, o investigador. Mas, por sua vez, quer o trabalho do utilizador interno, quer o do utilizador externo estão condicionados pela plataforma e,

¹⁴ Na página do TEITOK, em <http://www.teitok.org/index.php?action=projects> podem ser consultados alguns dos projetos de *corpora* desenvolvidos, ou em desenvolvimento: *corpora* orais, escritos, históricos, de referência e académico.

em última instância, pelo administrador que supervisiona todas as ações. Assim, descrever a plataforma implica descrever as suas potencialidades por parte de quem a utiliza em diferentes momentos. Adotaremos as duas perspectivas em função de cada uma destas valências, uma vez que, como se poderá constatar mais à frente neste trabalho (parte II), apesar de o ponto de partida para este projeto ter sido a perspectiva do utilizador externo, muito do trabalho desenvolvido só foi possível recorrendo às funcionalidades da plataforma na perspectiva do utilizador interno.

Por fim, e como o trabalho realizado com *corpora* em fases iniciais - armazenamento e organização dos dados e metadados - e fases intermédias - tokenização, lematização e anotação linguística - se traduz na disponibilização dos dados ao público, assume especial relevância a descrição do CQP (*Corpus Query Processor*) disponibilizado pela plataforma TEITOK para realizar pesquisas *online*.

2.1. A criação de um *corpus* informatizado

Como vimos anteriormente, a criação de um *corpus* de aprendentes passa inevitavelmente pelo seu carácter informatizado. Neste sentido, cabe ao criador do *corpus* seleccionar o *software* que melhor permite estruturar digitalmente o seu *corpus*, de acordo com os critérios externos e internos previamente estabelecidos. Por isso, quando se fala, aqui, em criação de *corpus*, subentende-se criação do *corpus* informatizado, isto é, conceção de uma estrutura digital que permita organizar e explorar o conjunto dos dados.

O TEITOK possibilita, ao utilizador interno, o *upload* de ficheiros de diferentes formatos, por exemplo, *XML*, *mp3*, *mp4*, *jpeg* e *PDF*. Este aspeto constitui uma enorme vantagem, na medida em que a plataforma permite armazenar diferentes tipos de dados, em diferentes níveis estruturais. É possível armazenar as produções dos aprendentes em ficheiros *XML* (transcrição de produções orais e escritas), ficheiros *mp3* (produções orais) e ficheiros *jpeg* (digitalização de textos escritos originais) e, simultaneamente, compilar os metadados que sustentam o *corpus* em formatos diversos (*XML*, *mp4* e *PDF*).

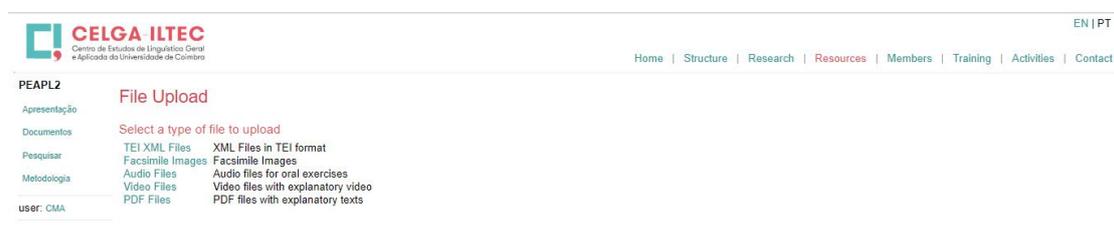


Figura 2: Interface do PEAPL2: opções de ficheiros suportados pela plataforma.

Neste domínio, os ficheiros *XML* são a peça-chave no ambiente TEITOK, pois «In TEITOK, a corpus consists of a collection of XML files, each in the Text Encoding Initiative (TEI) format (Janssen, 2016:4037). É o armazenamento da informação em ficheiros *XML*, após a sua conversão para o formato TEI, que permite a “preparação” e edição dos dados e dos metadados para pesquisa futura através de campos de pesquisa. A figura que se segue exemplifica os passos a realizar para criação de um ficheiro *XML*.

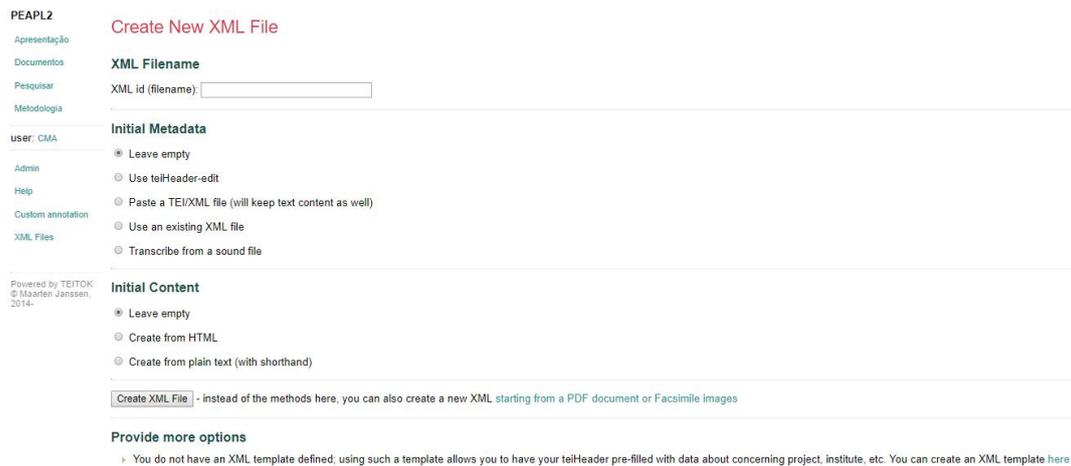


Figura 3: Interface do PEAPL2: criação de ficheiro XML.

No entanto, todas as informações disponibilizadas nos mais diversos formatos devem fazer parte do *corpus* e poder encontrar-se em diferentes locais, de acordo com o desenho estrutural do *corpus*. Para isso, a plataforma permite o acesso à informação através da criação de (sub)páginas e de (sub)secções, acessíveis a partir de menus personalizáveis, de acordo com as especificidades de cada *corpus*, onde é possível encontrar ficheiros de dados e descrições destes. Para o efeito, o utilizador interno pode criar páginas HTML, onde introduz diretamente informações, e que pode personalizar e editar, sempre que necessário. Observe-se a figura 4 a este propósito.

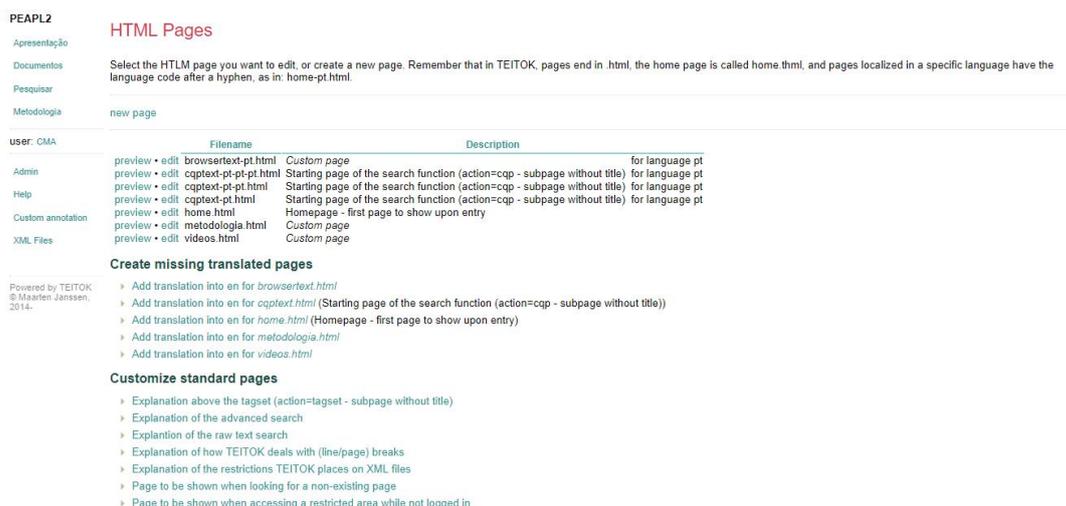


Figura 4: Interface do PEAPL2: criação e/ou edição de páginas HTML.

2.2. A preparação e edição de dados

Os dados que constituem um *corpus* são de natureza distinta: existem os dados respeitantes às produções dos aprendentes e os dados referentes aos metadados.

Relativamente aos dados textuais, a sua preparação passa por três etapas que se sucedem cronologicamente - *transcrição*, *tokenização* e *anotação* - que importa descrever com algum detalhe.

Transcrição

A transcrição das produções escritas e orais dos aprendentes pode ser feita diretamente na plataforma, a partir de uma página HTML ou a partir de um documento de escrita simples, sem formatações, por exemplo em *plaintext*, que será carregado posteriormente no sistema com as formatações de um documento *XML*.

Nesta fase de transcrição, deverão ser definidos critérios a adotar em função do tipo de anotação que será disponibilizada para o texto. No TEITOK, é possível definir convenções de transcrição para os *corpora* escritos e para os *corpora* orais. Na transcrição de textos escritos, é possível codificar segmentos que foram rasurados pelos aprendentes (legíveis e ilegíveis), segmentos acrescentados ou leituras conjeturadas. Este procedimento permite disponibilizar, futuramente, ao utilizador externo, diferentes representações do texto de acordo com as suas perguntas de investigação. Relativamente à transcrição de produções orais, também é possível dar conta de hesitações, reformulações, repetições, truncamentos e pausas preenchidas (segmentos paratextuais e extratextuais). Ao longo da transcrição de uma produção oral, o texto pode ser segmentado em unidades menores, respeitando as características prosódicas do texto, nomeadamente a pausa longa.

Embora todos estes segmentos que são transcritos mediante convenções definidas para cada *corpus* sejam visíveis no texto escrito / transcrito e audíveis na produção do falante, estes não são pesquisáveis no *corpus*.

Para os *corpora* orais, o TEITOK permite, ainda, alinhar a transcrição com os ficheiros áudio, possibilitando a visualização da onda sonora e a audição dos vários segmentos textuais ou do texto integral.



Figura 5: Interface do COral-Co: representação da onda sonora.

Tokenização

O sistema TEI - *Text Encoding Initiative* - permite segmentar os textos em unidades de significado a que se dá o nome de *tokens*, um sistema de codificação «where each token is assigned an integer number that represents its position in the corpus, starting at zero.» (Hardie, 2012:389). A tokenização é um processo automático, e, por isso, simples e rápido de executar pelo utilizador interno, uma vez que essa opção é disponibilizada, por pré-definição, pela plataforma e está ao alcance de um clique. Observe-se um mesmo texto antes e após a tokenização (figuras 6 e 7).

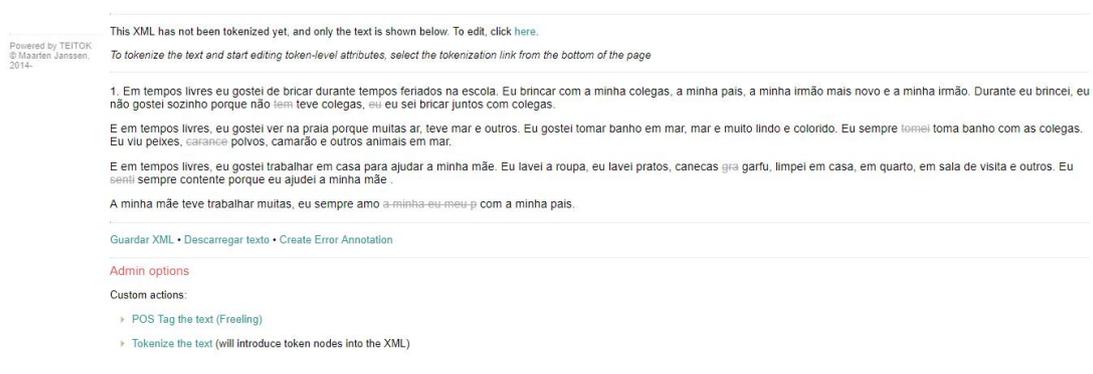


Figura 6: Interface do corpus de Timor: opção de tokenização.



Figura 7: *Interface do corpus de Timor: texto tokenizado.*

Este procedimento é essencial para todas as outras etapas que se seguem porque a tokenização cria «a straight-forward merge of the textual and the linguistic annotation» (Janssen, 2016:4038), visto que, ao segmentar as unidades, permite atribuir etiquetas de natureza diversa a cada uma delas. Este processo realiza-se a partir da transcrição do texto original do aprendente, seja escrito ou oral, e não a partir de versões corrigidas.

Recorde-se que, além das produções dos aprendentes, também os metadados que tiverem sido disponibilizados em formato *XML* são tokenizados, para serem codificados pelo sistema e, posteriormente, serem pesquisáveis através de campos criados para o efeito. O modelo de pesquisa utilizado por esta plataforma, e que a seguir descreveremos, vê o *corpus* «as consisting of a stream of tokens» (Hardie, 2012:389).

Anotação

O processo de anotação não é uma etapa obrigatória, podendo um *corpus* não ser anotado, como resultado de opções teóricas e da natureza das questões que este procedimento levanta¹⁵. No âmbito deste projeto, trabalharemos com *corpora* de aprendentes anotados e, como tal, justifica-se esclarecer de que forma este processo pode ser realizado, pelo utilizador interno, na plataforma TEITOK.

A anotação pode realizar-se a dois níveis: a) a anotação linguística incide sobre as unidades de significado que constituem o texto e b) a anotação textual visa as diferentes formas de representação do texto em função das “modificações” a que foi sujeito (por exemplo, transcrição, texto final do aprendente ou forma corrigida).

a) Anotação linguística

¹⁵ Algumas questões, bem como o conceito de anotação, já foram abordados no capítulo I, aquando da definição de *corpus* linguístico, pelo que aqui só se retomarão aquelas que possam ser úteis no âmbito da anotação de *corpora* de aprendentes de L2.

Terminado o processo de tokenização, o TEITOK possibilita a realização de um processo automático de anotação linguística¹⁶, numa primeira fase, a dois níveis: lema e categorias morfossintáticas.

A lematização consiste em etiquetar as palavras, aqui entendidas como unidades de significado, com a sua forma correspondente à entrada de dicionário. Este é um processo automático e que depende de como foram segmentadas as unidades de significado no momento da tokenização. O TEITOK reconhece cada uma das palavras simples e compostas e cada um dos sinais de pontuação como apenas um lema e considera dois lemas os casos das preposições contraídas e dos verbos na conjugação pronominal. Não são considerados lemas, no caso dos *corpora* orais, os segmentos extralinguísticos e paralinguísticos.

Ao nível da anotação de natureza morfossintática, o TEITOK permite atribuir uma classe/subclasse a cada uma das unidades, com informação de género gramatical, número e outras com relevância para a flexão. Para que isso seja possível, é preciso definir as categorias morfossintáticas relevantes para uma determinada língua-alvo e criar um sistema de etiquetas que será utilizado na categorização. Para que o sistema reconheça automaticamente as classes e subclasses a que pertence cada unidade, é utilizado «a definition file tagset.xml that explains all the different positions in the tagset» (TEITOK, *online help page*). A informação contida nas etiquetas é constituída por uma forma abreviada que, por sua vez, resulta das abreviaturas atribuídas às designações de classes e subclasses, de acordo com o exemplo que se segue.

¹⁶ A anotação linguística pode incidir sobre diversas estruturas da língua - morfossintática, semântica, fonológica, pragmática e lexical - como havia já sido retratado no capítulo I.

ID do token: w-130

venir

Tu vais ver quando ~~ta~~ tu vais venir visitar-me.

Classe morfofssintática	VMN0000
Lema	vir

Etiqueta: VMN0000

V	POS principal	Verb
M	Type	main
N	Mood	infinitive
0	Tense	não se aplica
0	Person	não se aplica
0	Num	não se aplica
0	Gen	não se aplica

Figura 8: Interface do TEITOK: lema e classificação morfofssintática de um *token*.

Embora seja automático e poupe muito tempo e trabalho, este processo não dispensa a verificação manual, pois há casos em que o *software* não consegue desfazer a ambiguidade de determinadas estruturas linguísticas, cabendo ao anotador esse trabalho *a posteriori*. Observe-se, a este propósito, o exemplo que se segue, em que a palavra *jantar* foi etiquetada como verbo e não como nome, uma vez que o sistema não faz a distinção do contexto.

Ontem Adv ontem	fui Verbo ir	ão Prep+Det a+o	jantar Verbo jantar	de Prep de	curso Nome curso	com Prep com	todos Det todo	os Det o	colegas Nome collega	da Prep+Det de+o	universidade Nome universidade	:	foi Verbo ir	giro Adj giro	! Pontuação !
-----------------------	--------------------	-----------------------	----------------------------------	------------------	------------------------	--------------------	----------------------	----------------	----------------------------	------------------------	--------------------------------------	---	--------------------	---------------------	---------------------

Agora Adv agora	tenho Verbo ter	de Prep de	jantar Classe morfofssintática Verbo (VMN0000) principal; infinitivo Lema jantar		:
-----------------------	-----------------------	------------------	--	--	---

Figura 9: Interface do PLE: erro de etiquetagem no processo automático de anotação.

Numa segunda fase, a partir da lematização e da classificação morfofssintática, é possível realizar uma anotação com base em erros cometidos pelo aprendente, ou pelo menos numa suposição «on what the student should have written» (Janssen, 2019). O TEITOK permite classificar e corrigir erros de natureza ortográfica, sintática e lexical. No entanto, observam-se algumas questões técnicas que podem condicionar este nível de anotação:

a. Uma vez que anotação é baseada em *tokens*, torna-se muito difícil anotar unidades segmentais menores (por exemplo, ao nível do grafema ou da sílaba) ou maiores (estruturas que constituem unidades multilexicais, expressões);

b. A anotação nos *corpora* orais segue o mesmo procedimento, baseado nos textos escritos resultantes da transcrição das produções orais, não havendo lugar para a anotação linguística relacionada com as características do discurso oral;

c. Face às limitações apresentadas em a. e b., prevê-se a ausência de anotação linguística em determinados níveis e, conseqüentemente, a criação de sistemas de anotação em função de novos níveis de anotação linguística de forma a poder automatizar o procedimento.

O Cople2 é o único *corpus* português com anotação de erros desenvolvida com base no sistema TEI, utilizado pelo TEITOK. Para levar a cabo esta tarefa, os seus criadores desenvolveram um «fine-grained tag set system» (Río & Mendes, 2018:23) para complementar a anotação linguística realizada a partir da tokenização. Apesar dos progressos já alcançados no processo de anotação de erros, as autoras afirmam, num artigo intitulado *Error annotation in the COPLE2 corpus* (Río & Mendes, 2018), que há, ainda, um longo caminho a percorrer.

First of all, we need to explore how to transform the multi-token in-line annotations into tags, reducing as much as possible the manual effort. (...) A second line of work is related to the addition of new linguistic areas for error annotation, like semantics or discourse (Río & Mendes, 2018:236).

Isto significa que a anotação de natureza semântica, fonológica, pragmática e lexical, embora possível no TEITOK, não existe “fora” da anotação de erros, isto é, é com base no erro que se faz a anotação linguística, quando o desejável seria o processo inverso: primeiro anotar e depois corrigir. É preciso, então, ter em conta que se trata de processos distintos de anotação linguística: estabelecer diferentes níveis de anotação e proceder à anotação de erros relativos a cada nível.

b) Anotação textual

Para além das etiquetas de carácter linguístico, o texto pode apresentar outras ao nível da representação do texto, permitindo visualizar os textos de acordo com as opções de transcrição definidas. Assim, a transcrição contempla duas representações: o texto original, com todas as

hesitações, reformulações e adições, e o texto final do aprendente, “limpo” de todos estes traços que envolvem o processo de escrita. No que concerne aos *corpora* orais, para além das diferentes representações do texto, é ainda possível ouvir os ficheiros áudio, que reproduzem na totalidade, ou de forma segmentada, as produções dos aprendentes, e observar a representação da onda sonora.

Opções de representação

Texto: Transcrição Forma do aluno Forma corrigida - Mostrar: Cores - Etiquetas: Classe morfosintática Lema

Figura 10: *Interface* do TEITOK: opções de representação do texto.

Uma etapa igualmente importante nesta fase de preparação dos dados é a de organização dos metadados.

Os metadados estão organizados em dois níveis distintos, ao nível do aprendente e ao nível dos critérios externos e internos de conceção do *corpus*. Os primeiros podem encontrar-se no cabeçalho (*TeiHeader*), que acompanha a produção do aprendente, ou em secção à parte, mas facilmente pesquisável. Nestes casos, como a informação é processada em formato *XML*, esta pode ser pesquisada posteriormente, através de campos de pesquisa. No que se refere ao segundo grupo de metadados, estes podem figurar em páginas HTML (onde se apresenta o *corpus*, a metodologia, o protocolo de recolha, por exemplo) que, por sua vez, podem conter remissões para outras páginas, secções ou documentos noutros formatos. Nesta situação, a plataforma permite navegar no *corpus* acedendo facilmente a esta informação.

Todo o trabalho resultante de preparação de dados, descrito até ao momento, especialmente o de anotação linguística, pode, a qualquer momento, ser revisto e editado. O TEITOK permite a edição de ficheiros *XML* que contêm os dados que constituem o *corpus*, sempre que se justifique acrescentar ou corrigir alguma informação. Para agilizar este processo de edição, ao nível da anotação linguística, por exemplo, quando é detetada uma falha ao nível da classificação morfológica, o TEITOK facilita esta tarefa ao permitir encontrar, através da área de pesquisa, uma determinada palavra, ou classe de palavras e, ao invés de as alterar uma a uma, seleccionar todas as visadas e proceder à alteração.

PEAPL2

Apresentação

Documentos

Pesquisar

Metodologia

user: CMA

Admin

Help

Custom annotation

XML Files

Powered by TEITOK
© Maarten Janssen, 2014.

Multiple token edit via CQP Search

Define below which features you want to change in this search, and select all the tokens for which you want that change to be made. Leaving a feature empty will not eliminate it's value, but just ignore that feature in the edit.

The CQP corpus can become disaligned wrt the XML files after editing tokens. Therefore, always regenerate the CQP corpus before using this function!

Click here to enter individual values for each result

626 resultados for [form="casa"] • A mostrar 0 - 500 (seguintes)

File ID	Set.	Left context	Match	Right context
xmlfiles/timor/MAM SEC.12.DS.04.1.xml	<input type="checkbox"/>	praia ou	casa	outro amigo.
xmlfiles/timor/NR SEC.12.DS.10.5.xml	<input type="checkbox"/>	uma tia	casa	a minha tia todos os
xmlfiles/timor/NR SEC.12.DS.10.5.xml	<input type="checkbox"/>	de manha . A minha	casa	é cor de rosa tem
xmlfiles/timor/NR SEC.12.DS.10.5.xml	<input type="checkbox"/>	tenho três televisão	casa	Uma sala
xmlfiles/timor/NR SEC.12.DS.10.5.xml	<input type="checkbox"/>	meu pai . Tem duas	casa	de banho . Eu gosto
xmlfiles/timor/NR.BAS.09.NA.18.1.xml	<input type="checkbox"/>	missa vou para	casa	11.30 horas e
xmlfiles/timor/NR.BAS.09.NA.18.1.xml	<input type="checkbox"/>	e faz o trabalho de	casa	que os professores damos e
xmlfiles/timor/TET SEC.12.DS.02.1.xml	<input type="checkbox"/>	ajudar os meus pais em	casa	. Eu ajudar os meus
xmlfiles/timor/TET SEC.12.DS.02.1.xml	<input type="checkbox"/>	ajudar os meus pais em	casa	eu lavra as roupas ,
xmlfiles/timor/TET.BAS.09.NA.22.5.xml	<input type="checkbox"/>	xxx . dona de	casa	. A minha irmã mais

Figura 11: Interface do TEITOK: opções de edição de múltiplos dados.

Já quando se pretende corrigir apenas algumas formas, esta tarefa pode realizar-se caso a caso, preenchendo os campos pretendidos, como se pode observar no exemplo:

PEAPL2 PLE

Apresentação

Documentos

Pesquisar

Metodologia

user: CMA

Admin

Help

Custom annotation

XML Files

Powered by TEITOK
© Maarten Janssen, 2014.

Edit Token

Filename | romeno a1.14.6.1.b.xml

Title | Sem título

Token value (w-130): venir

user: CMA

Admin

Help

Custom annotation

XML Files

Inserted after: attached / separate • before: attached / separate • insert elm before: paragraph ; linebreak • split in dtoks: 2 ; 3

edit context XML • merge left to w-129 • create mtok left: 1 ; 2

treat similar tokens

Tu vais ver quando tu vais venir visitar-me.

Save Cancel • Token Details

Figura 12: Interface do TEITOK: opções de edição de dados.

A cada edição de dados, o TEITOK procede à regeneração automática do *corpus*, para que o sistema assuma as alterações efetuadas.

O trabalho de criação do *corpus* e de preparação e edição dos dados é da responsabilidade do utilizador interno da plataforma, ainda que muitas vezes condicionado pelas características da plataforma TEITOK. Destaca-se, face ao exposto, a versatilidade dos ficheiros XML, que permitem armazenar e editar informação e torná-la pesquisável na plataforma.

2.3. Pesquisa

A pesquisa *online* é um instrumento valioso para a exploração e extração de informação de *corpora* de aprendentes. A área de pesquisa e as suas funcionalidades são concebidas pelo administrador da plataforma e «the core functions of TEITOK is to make your corpus searchable by creating an indexed corpus from the XML files using the Corpus Workbench/Corpus Query Processor (CQP)» (Janssen, 2014). Vejamos, então, de uma forma muito simplificada, como funciona o CQP.

Tal como já havíamos referido, toda a informação contida em ficheiros XML é pesquisável. Veja-se, como exemplo, os diferentes níveis de informação, listados em colunas, num ficheiro XML relativo à produção de um aprendente.

Verticalized Corpus View

XML File: ple/bulgaro.a2.31.1.1a.xml

	Transcription	Student form	Orthographically corrected form	POS tag	Lemma
w-1	Sou			VMIP1S0	ser
w-2	um			Dl0MS0	um
w-3	Búlgaro		búlgaro	AQ0MS00	búlgaro
w-4	e			CC	e
w-5	eu			PP1CSN0	eu
w-6	vivia			VMI3S0	viver
w-7	em			SP	em
w-8	Portugal			NP00000	Portugal
w-9	4			Z	4
w-10	meses			NCMP000	mês
w-11	.			Fp	.
w-12	Era			VMI1S0	ser
w-13	um			Dl0MS0	um
w-14	grande			AQ0CS00	grande
w-15	prazer			NCMS000	prazer
w-16	por			SP	por
w-17	mim			PP1CSO0	mim
w-18	.			Fp	.
w-19	Estou			VMIP1S0	estar
w-20	longo			AQ0MS00	longo
w-21	.			Fc	.
w-22	com			SP	com
w-23	cabelo			NCMS000	cabelo
w-24	preto			AQ0MS00	preto
w-25	e			CC	e

Figura 13: Interface do TEITOK: opções de edição de dados.

Cada uma destas colunas de informação é pesquisável através de campos de pesquisa. Tal significa que a pesquisa é o reflexo do trabalho de preparação dos dados e, por isso, os dois, preparação e pesquisa de dados, devem estar em consonância.

De acordo com as particularidades de cada *corpus* e com os níveis de anotação linguística definidos, é possível que o CQP não os assuma automaticamente nos seus campos de pesquisa. «Instead, the administrator can configure CEQL on a per-corpus basis to link any available annotation» (Hardie, 2012:399).

Cabe, ainda, ao administrador da plataforma levar a cabo a formatação do aspeto gráfico do *interface*, que, no caso da plataforma TEITOK, se apresenta da seguinte forma ao utilizador externo:

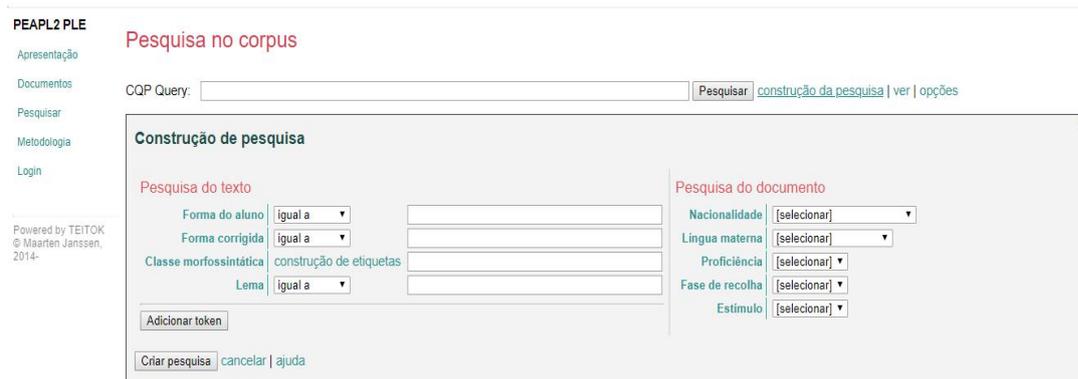


Figura 14: Interface do TEITOK: aspeto gráfico da área de pesquisa do PEAPL2.

E para o utilizador externo, quais são as funcionalidades e as vantagens da pesquisa de *corpora* através da plataforma TEITOK utilizando o CQP?

Como pesquisar

A pesquisa no *corpus* pode ser feita por uma simples palavra, ou expressão, e/ou preenchendo os campos de pesquisa disponíveis, que podem ser referentes a (i) etiquetas de natureza linguística disponíveis para cada um dos textos e (ii) metadados que integram o cabeçalho. A área de pesquisa, quando personalizada de acordo com a organização dos dados que é possível extrair do *corpus*, é, na sua generalidade, intuitiva.

Gostaríamos de destacar, neste ponto, a pesquisa por lema, por a considerarmos crucial na pesquisa de *corpora* de aprendentes. Quando os aprendentes produzem enunciados não coincidentes com a língua-alvo, não é possível encontrar as formas desviantes apenas com base na intuição de falante nativo, embora se possa circunscrever, em casos específicos, as áreas em que estas possam ocorrer. Nestes casos, a pesquisa por lema permite realizar uma pesquisa por unidade de significado, cujos resultados são todas as formas que com ela se relacionam (variação em número, género, flexão verbal, formas desviantes).

No entanto, algumas dificuldades poderão surgir quando se pretende pesquisar, por exemplo, estruturas linguísticas mais complexas. É possível fazê-lo utilizando os comandos *Adicionar token* ou *A acrescentar ao lado*, sendo que esta funcionalidade, para já, apenas está disponível na área de pesquisa dos *corpora* do CELGA-ILTEC-Coimbra.

Adicionar token permite pesquisar uma sequência de *tokens*, isto é, adicionar uma sequência de unidades que se podem referir à *forma do aluno*, à *forma corrigida* ou ao *lema*, mas esta opção não se pode aplicar à pesquisa por *classe morfossintática*. Já a opção *Acrescentar ao lado* permite pesquisar classes morfossintáticas em contextos sequenciais simples a partir de uma classe definida, mas alternativos, isto é, pesquisar uma sequência do tipo a+b **ou** a+c, mas não do tipo a+b+c, como exemplificaremos mais à frente.

Quando as opções pré-definidas na área de pesquisa não são suficientes para aceder à informação que se procura, então, nestes casos, a solução passa por realizar uma pesquisa manual avançada, usando expressões de pesquisa que, embora não sendo complexas, implicam o conhecimento da linguagem de pesquisa inerente ao CQP na plataforma TEITOK.

Assim, a pesquisa de segmentos linguísticos deve respeitar a seguinte estrutura:

[etiqueta="xxx"]

sendo que os códigos das etiquetas correspondem a formas abreviadas, na língua inglesa, com a seguinte equivalência:

pos = classe morfossintática

lemma = lema

form = forma do aluno

nform = forma corrigida (PEAPL2_PLE)

word = forma corrigida (COraL-Co)

Supondo que pretendemos pesquisar qualquer forma do verbo *fazer*, deveremos digitar a pesquisa nos seguintes moldes:

[lemma="fazer"]

A expressão de pesquisa é igual para a pesquisa por *Forma do aluno* e por *Forma corrigida*, variando apenas a informação a pesquisar. Mas, quando queremos pesquisar por *Classe morfossintática*, a esta estrutura devemos ainda acrescentar os seguintes símbolos à expressão:

[etiqueta="xxx.*"]

Para além disso, é preciso conhecer todas as abreviaturas respeitantes às diferentes classes e subclasses morfossintáticas¹⁷ que figuram no *corpus* para poder completar a expressão de pesquisa. Por exemplo, se pretendermos pesquisar todos os adjetivos qualificativos, no masculino, singular, existentes no *corpus*, devemos formular a expressão da seguinte forma:

[pos="AQ.MS*"]

Podemos combinar várias expressões de pesquisa de modo a extrair a informação pretendida. Por exemplo, em função do contexto de ocorrência, podemos fazer uma pesquisa do tipo: adjetivo qualificativo seguido de nome:

[pos="AQ.*"] [pos="N.*"]

Podemos, ainda, querer saber qual o contexto mais frequente de ocorrência de determinadas classes/subclasses em função da sua colocação. Por exemplo:

[lemma="fazer"] [pos="N.*"|"RG.*"]

Neste caso, a pesquisa pretende encontrar resultados para o verbo *fazer* seguido de nome comum ou, em alternativa (condição aqui representada por barra vertical |), de advérbio, como se pode observar na listagem de resultados obtidos com esta expressão de pesquisa:

¹⁷ O documento contendo o sistema de etiquetas disponíveis para as classes/subclasses de palavras será disponibilizado em anexo (anexo 1).

PEAPL2 PLE

Apresentação

Documentos

Pesquisar

Metodologia

Login

Pesquisa no corpus

CQP Query: [lemma="fazer"] [pos = "NC |R.*"] [Pesquisar] construção da pesquisa | ver | opções

290 resultados • A mostrar 0 - 100 (seguintes)

Texto: [Transcrição] [Forma do aluno] [Forma corrigida]

Etiquetas: [Classe morfosintática] [Lema]

Powered by TEITOK
© Maarten Janssen,
2014.

contexto	que seja mais sadável a	fazerexerciso	Também, eu adoro
contexto	CAMIÕES, TRACTORES COMBOIOS E	FAZEM COMPETIÇÕES	[...]MESMO INGRASSADO!!
contexto	estou indo para Natal.	Fazemos alguns	biscoitos junto para Natal em
contexto	banho de sol .,	Fazemos castelos	[...] das areias, ficou
contexto	só gosta de caminhar.	Fazer desporto	é importante para sua saúde
contexto	gosto de fazer desporto.	Faço jogging	varias vezes a semana.
contexto	ano no Brasil.	Fiz intercâmbio	lá, isto significa que
contexto	ir estudar a Granada.	Fizeste bem	. Vou tentar visitar-te
contexto	tenho a certeza que	farei bem	nos meus exames.
contexto	uma experiência como esta te	faria muito	bem. Antes de tudo
contexto	voltar a falar contigo como	fásiamos Antes	quando estavam mas perto de
contexto	(cunhas curtias)	faz amizade	e com outras não conectas
contexto	tempo quente já que aqui	faz bastante	frio. (Sei que
contexto	si. Normalmente a natureza	faz bem	às pessoas e é
contexto	nova. Acho que a variabilidade	faz bem	e que se todos fossemos
contexto	O sul de país sempre	faz calor	no oeste é
contexto	com tempo melhor porque sempre	faz chuva	por isso não posso fazer
contexto	irmã tem 14 anos e [...]	faz escola	secundaria. Ela joga e
contexto	Coimbra mas o que me	faz falta	é o meu carro,
contexto	Conta-me tudo,	faz favor	, e não me deixes
contexto	cama também, quando	faz frio	ou chova no exterior
contexto	Europa. Bulgária	faz fronteira	com o Mar Negro ao
contexto	parte do ano quando	faz frio	, costume ficar muito tempo
contexto	talvez no verão quando	faz mais	sol e calor porque agora

Figura 14: Listagem dos resultados obtidos para a expressão de pesquisa avançada, realizada no PEAPL2.

Quando a expressão de pesquisa é constituída por vários elementos, estes são pesquisados pela ordem em que surgem na expressão. A pesquisa avançada pode conjugar-se com a seleção de informação dos campos de pesquisa referentes aos metadados.

Para os *corpora* orais, o funcionamento da pesquisa é bastante semelhante (apenas alguns campos de pesquisa podem ser personalizados de acordo com a natureza do *corpus*), uma vez que a pesquisa se baseia nos textos escritos resultantes da transcrição das produções orais, não havendo lugar à pesquisa diretamente a partir de segmentos orais.

Antes de se realizar uma pesquisa, deve ter-se em conta quais os critérios subjacentes à segmentação de unidades de significado e à lematização do *corpus*, pois os resultados apresentados dependerão destas opções e poderão condicionar a leitura dos dados. É essencial consultar, portanto, a informação disponibilizada pelo *corpus*.

Os resultados de pesquisa

Os resultados da pesquisa são apresentados em linha de contexto KWIC (*Key Words In Context*) podendo ser visualizados num contexto reduzido (5 *tokens*) ou num contexto mais alargado (até 30 *tokens*), como pode ser observado nas figuras 15 e 16.

PEAPL2_PLE

Apresentação

Documentos

Pesquisar

Metodologia

Login

Powered by TEITOK
© Maarten Janssen,
2014.

Pesquisa no corpus

CQP Query: [construção da pesquisa](#) | [ver](#) | [opções](#)

Opções de busca

Tipo de representação visual: KWIC Context

Tamanho do contexto: palavras

Ordenar por:

Estratégia de combinação:

71 resultados

Texto: [Transcrição](#) | [Forma do aluno](#) | [Forma corrigida](#)

Etiquetas: [Classe morfosintática](#) | [Lema](#)

contexto	lá, em Inglaterra,	janta-se	às sete horas
contexto	de pato. Depois de	Jantar	, fomos à casa
contexto	para nadar, passar ou	Jantar	pequena pela praia com
contexto	menos em Coimbra, se	janta	mais tarde, pelas
contexto	. Normalmente, os japoneses	jantam	às 19 a 20
contexto	viajar. Nós almoçamos e	jantamos	num muito bonito restaurante
contexto	Toda a família nos	jantamos	para intercambiar prendas, e
contexto	cinema com meus amigos e	jantamos	fora.
contexto	noite cm meus amigos,	jantar	fora e ir ao
contexto	. Ontem fui ao	jantar	de curso com todos os
contexto	almoço ou no	jantar	. Os portugueses gostam de
contexto ao restaurante e	jantar	fora cum os meus pais
contexto	Meu amigo espanhol começa a	jantar	às 21 ou 22.

Figura 15: Resultados de pesquisa apresentados em linha de contexto (5 tokens) - PEAPL2_PLE.

Pesquisa no corpus

CQP Query: [construção da pesquisa](#) | [ver](#) | [opções](#)

Opções de busca

Tipo de representação visual: KWIC Context

Mostrar contexto: Tokens

Ordenar por:

Estratégia de combinação:

71 resultados

Texto: [Transcrição](#) | [Forma do aluno](#) | [Forma corrigida](#)

Etiquetas: [Classe morfosintática](#) | [Lema](#)

contexto ano estou a estudar em Portugal. Sendo Inglesa, consigo notar algumas diferenças culturais. Para começar, o horário é completamente diferente; lá, em Inglaterra, **janta-se** às sete horas da tarde, enquanto cá, em Portugal, ou pelo menos em Coimbra, se janta mais tarde, pelas nove

contexto | | Eu estava com fome, por isso, logo que nos encontramos no Rodobiário, fomos jantar juntos, comemos arroz de pato. | Depois de **Jantar**, fomos à casa dela de autocarro e bebemos vinho tinto para festejarmos o aniversário dela. Mas eu não tinha tempo. Naquele dia,

contexto | | Também, nos visitamos nossa casa de praia para mudar e usamos nosso tempo livre u pouco diferente que normal. Vamos a praia para nadar, passar ou **Jantar** pequena pela praia com cerveja. Nossa casa de praia é muito tranqul, por isso, eu | | gostou de ler ou ouvir musicas ali. Muito pacifico.

contexto | | lá, em Inglaterra, janta-se às sete horas da tarde, enquanto cá, em Portugal, ou pelo menos em Coimbra, se **janta** mais tarde, pelas nove horas. | Também já reparei que há uma diferença quanto a vida nocturna. Lembrei-me da primeira vez que sai com

contexto rara, por isso não tinha me divertido no inicio, mas agora já costumei. | Também é muito diferente como passar o dia. Normalmente, os japoneses **jantam** às 19 a 20, mas, no caso dos portugueses, é mais tarde. | Meu amigo espanhol começa a jantar às 21 ou 22

contexto | | meus sapatos. Um dia nos fomos em Birro. Birro foi muito bonito. Nos tivemos o carro, então nos podemos viajar. Nos almoçamos e **jantamos** num muito bonito restaurante. | Já sou em Portugal. Eu estudo na universidade de Coimbra. E eu sou diferente. Eu sou magra e tenho o

contexto também é muito importante para mim, e passo muito tempo com ela. Gosto muito dos jantares familiares, sobretudo dos aniversarios. Toda a família nos **jantamos** para intercambiar prendas, e é muito jiro. Da minha familia, a minha irmã pequena é um pilar muito importante. Temos uma idade parezida pelo

Figura 16: Resultados de pesquisa apresentados em contexto (30 tokens) - PEAPL2_PLE.

A partir dos resultados extraídos de uma pesquisa, é possível observar os dados de acordo com determinadas opções de frequência e armazenar essa informação em ficheiros *excel* para posterior tratamento. A plataforma também possibilita que sejam guardadas, temporariamente, algumas expressões de pesquisa, criando um histórico, através do qual se pode proceder à comparação dos dados resultantes de cada pesquisa efetuada.



Figura 17: Opções disponibilizadas pelo TEITOK relativamente à lista de resultados de pesquisa: “Descarregar resultados”, “Memorizar expressões de busca” e “Opções de frequência”.

Ao consultar cada um dos textos, também é possível consultar as respetivas anotações linguísticas, bem como a forma original do texto (seja transcrita, digitalizada ou em *facsimile*) e, ainda, aceder aos dados do cabeçalho (simples ou expandido)¹⁸.



Figura 18: Exemplo de uma produção escrita extraída do PEAPL2_PLE: cabeçalho e opções de representação do texto.

¹⁸ Em alguns casos, os dados sobre o texto podem estar numa secção distinta, mas igualmente pesquisável.

No caso dos *corpora* orais, ao selecionar o contexto, tem-se acesso não só à transcrição das produções orais (ou do texto/palavras para leitura), mas também ao ficheiro áudio correspondente. Deste modo, « TEITOK provides the unique option to have really mixed corpora with both written and spoken data, which is a crucial feature for learner corpora» (Janssen, 2019).

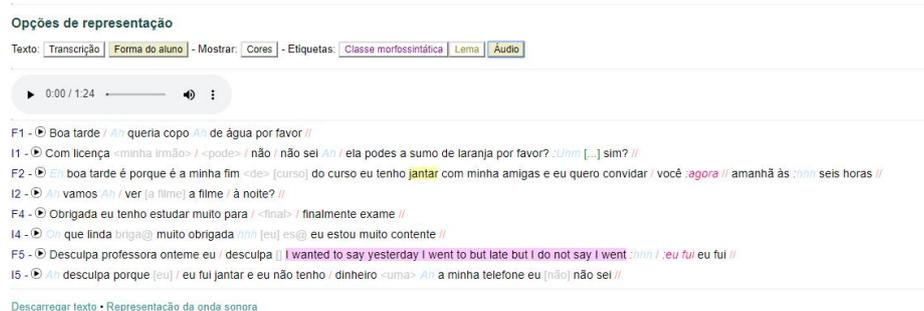


Figura 19: Exemplo de uma produção oral transcrita extraída do COraL-Co: opção *Áudio* (ficheiro integral ou segmentado).

O TEITOK permite, ainda, ao utilizador externo, armazenar os dados obtidos em diferentes formatos: a listagem de resultados da pesquisa em *text* ou *excel*, os textos escritos, em *text*, e os ficheiros áudio, em *mp3*.

Face ao que aqui fica exposto, no âmbito da pesquisa de dados através do CQP disponibilizado pelo TEITOK, reconhecem-se pontos fortes, mas também algumas fragilidades.

Sem dúvida, «CQPweb's main innovative feature is its flexibility, in that it is compatible with any corpus, rather than being bound to a particular dataset» (Hardie, 2012:381). Esta característica permite realizar pesquisas à medida de cada *corpus*, isto é, configurar a pesquisa em função da informação armazenada e da forma como esta está organizada em ficheiros *XML*, visando a consistência dos dados e a sua viabilidade em trabalhos com base na pesquisa de *corpora*.

Outra mais valia desta ferramenta de pesquisa é a possibilidade que oferece de observar os dados em contexto e por opções de frequência, permitindo a sua exportação para futura análise.

Cabe ainda referir que o TEITOK disponibiliza esta ferramenta de pesquisa através do *interface* da plataforma, atualizando os campos de pesquisa e personalizando a área de pesquisa (textos informativos, língua de pesquisa, formatação de fontes ao nível da cor e tipo de letra).

Por outro lado, e como refere Hardie (2012), tudo isto só é possível «since it implements the friendly and accessible user-interface design» (Hardie, 2012:381). Efetivamente, a descrição das expressões de pesquisa avançada que aqui apresentámos aponta para o facto de a linguagem de pesquisa nem sempre ser intuitiva e constituir um dos constrangimentos que o CQP apresenta, pois, apesar de a linguagem ser lógica, «even advanced users find difficult to memorize in all its details» (Hoffmann & Evert, 2006:180). Hardie (2012) reflete sobre os impactos que este constrangimento pode causar na pesquisa de *corpora* e chega à conclusão de que este fator pode influenciar as perguntas de pesquisa e, conseqüentemente, a consistência e a validade dos dados a extrair da pesquisa. Na verdade, o autor constata, em última análise, o seguinte:

It should be clear that given a choice between (i) learning to program or (ii) not using corpus data, the majority of potential corpus analysts will — understandably — opt for (ii) (Hardie, 2012:383).

De acordo com o autor, para ultrapassar esta situação, e visto esta ferramenta ser de grande utilidade não só para investigadores/linguistas, mas também para um público não especializado, «better online help will be an absolute necessity», aliás «help pages would be dynamically generated so that their content can be customized to the active corpus» (Hardie, 2012:405).

No caso do TEITOK, é disponibilizada uma página de *Ajuda* para o utilizador interno, no que se refere à criação, edição, visualização e pesquisa de *corpora* (embora contenha informações sobretudo de carácter descritivo, mais do que operacional), mas para o utilizador externo, não há qualquer secção de *Ajuda* e, tal como em grande parte dos *corpora* de aprendentes, «The online user documentation (“help pages”) in CQPweb is still very limited. (Hardie, 2012:405). É neste contexto que ganha particular relevância o presente projeto.

CAPÍTULO 3 - Dois *corpora* de aprendentes de PL2

INTRODUÇÃO

O PEAPL2_PLE e o COral-Co são dois *corpora* de aprendentes de português L2, de consulta livre, que visam a sustentação empírica da investigação na área de português língua não materna através das suas produções anotadas e pesquisáveis através da plataforma TEITOK.

Para além disso, são projetos afins que partilham a mesma metodologia no que se refere à seleção e à recolha de dados dos aprendentes. Provenientes de diversos países com diferentes línguas maternas, estes encontravam-se, à data de recolha dos dados, a estudar em Portugal, em cursos ou unidades curriculares de PLE da Faculdade de Letras da Universidade de Coimbra¹⁹, distribuídos por níveis de proficiência entre o A1 e o C1. Relativamente às produções recolhidas, é a sua natureza distinta, mas complementar, que distingue estes dois *corpora* de aprendentes de PL2: o PEAPL2_PLE enquanto *corpus* de produções escritas e o COral-Co enquanto *corpus* de produções orais. Esta distinção traduz-se, essencialmente, na preparação dos dados, mais concretamente no que se refere ao processo de transcrição, dadas as particularidades dos dois tipos de produções.

Estão ambos disponíveis para pesquisa *online* através da plataforma TEITOK e, por isso, partilham também alguns aspetos comuns quanto ao funcionamento da pesquisa de dados e metadados e ao *interface* da área de pesquisa.

Apesar do que os distingue, estes dois *corpora* assentam nos mesmos princípios estruturais e metodológicos, possibilitando uma análise comparativa dos dados.

¹⁹ Curso Anual, ou de Férias, de Língua e Cultura Portuguesas para Estrangeiros e unidades curriculares de Língua Portuguesa para *Erasmus*.

3.1. *Corpus* de produções escritas de aprendentes de PL2: subcorpus de português língua estrangeira (PEAPL2_PLE)

O subcorpus de Português Língua Estrangeira (PLE) integra o *Corpus* de Produções de Aprendentes de PL2 (PEAPL2), resultando de um projeto de Recolha de *Corpora* de PL2 iniciado em 2008, no Centro de Estudos de Linguística Geral e Aplicada (CELGA), com o objetivo de disponibilizar «um acervo estruturado de dados empíricos fiáveis, capazes de sustentar o desenvolvimento de dissertações na área da aquisição/aprendizagem de PL2». O PEAPL2 integra, para além do subcorpus PLE, o subcorpus Timor, já disponível publicamente, e o subcorpus Guiné-Bissau, ainda em fase de preparação.

O PEAPL2_PLE é constituído por 623 produções escritas por 458 aprendentes, de 39 línguas maternas diferentes e distribuídos por diferentes níveis de proficiência (entre o A1 e o C1), a estudar na Faculdade de Letras da Universidade de Coimbra, em Cursos de Português para Estrangeiros, entre os anos de 2009 e 2011, combinando, assim, situações de aprendizagem formal e em contexto de imersão. Para caracterização do perfil dos informantes foi elaborada uma ficha para preenchimento, com vista à recolha de informações relativas a dados pessoais e a dados sobre o seu percurso escolar, no que se refere à aquisição/aprendizagem da língua-materna e de outras línguas não maternas, entre as quais o português.

Os textos escritos produzidos pelos aprendentes foram obtidos a partir de um conjunto de nove estímulos que se dividem por três áreas temáticas do *Português Fundamental* (Nascimento *et al.*, 1987): *o indivíduo*, *a sociedade* e *o meio ambiente*. Estes estímulos foram extraídos da lista dos inicialmente propostos no projeto de *Recolha de dados de aprendizagem de português língua estrangeira*, coordenado por Isabel Leiria (Leiria, 2008). A cada um dos estímulos, descritos na Metodologia do projeto²⁰, corresponde um código para mais fácil categorização dos textos, mas também utilizado para a pesquisa de dados.

Posteriormente à recolha das produções escritas, estas foram transcritas seguindo as convenções de transcrição de Leiria (Leiria, 2006), por forma a preservar o texto original do aprendente e, por inerência, as informações que aquele pode aportar no âmbito da investigação na aquisição/aprendizagem de PL2. Assim, e como se poderá observar na tabela abaixo apresentada, as convenções a considerar no momento de consulta das produções identificam segmentos riscados legíveis, segmentos riscados ilegíveis, segmentos acrescentados, segmentos conjeturados e

²⁰ O protocolo de recolha dos dados pode ser consultado com maior detalhe em <http://teitok2.iltec.pt/peapl2-ple/index.php?action=metodologia>.

asseguram o anonimato no que se refere a nomes e/ou situações passíveis de identificar o informante. A representação gráfica das convenções adotadas aquando da fase de transcrição em formato *word*, foi sujeita a alguns reajustes quando os ficheiros foram transferidos para o ambiente TEI, em formato XML.

Word	XML
< (...) > segmentos riscados ilegíveis	{...} segmentos riscados ilegíveis
< xxx > segmentos riscados	___ segmentos riscados
/ xxx / segmentos acrescentados	Segmentos acrescentados (cor)
xxxx nomes próprios que permitam identificar o informante.	xxxx nomes próprios que permitam identificar o informante.
/* xxx / leituras conjeturadas	

Tabela 1: Convenções de transcrição utilizadas no PEAPL2_PLE, com base em Leiria (2006).

Os textos foram codificados em função dos seguintes elementos: i) língua materna do informante, ii) tipo de curso de PL2 frequentado pelo informante²¹, iii) nível da turma frequentada pelo informante, iv) nº de identificação do aprendente e v) código do estímulo usado para a produção do texto. Veja-se o exemplo que se segue:

Um texto sobre atividades dos tempos livres (33.1J) produzido pelo informante (01), de LM alemã (ALEMÃO), que frequentava, aquando da recolha de dados, uma turma de nível A2 destinada a alunos do programa *Erasmus* (ER) é identificável através do código: ALEMÃO.A2.01.33.1J²².

²¹ Na metodologia, apresentam-se os códigos atribuídos a cada curso da Faculdade de Letras da Universidade de Coimbra frequentado pelos aprendentes: ER (Erasmus), CA (Curso Anual de Língua e Cultura Portuguesas para Estrangeiros), CF (Curso de Férias de Língua e Cultura Portuguesas para Estrangeiros).

²² Esta informação pode ser consultada na metodologia do subcorpus PEAPL2_PLE, em <http://teitok2.iletec.pt/peapl2-ple/index.php?action=metodologia>.

Numa primeira fase, as produções escritas que agora integram o *PEAPL2_PLE* foram disponibilizadas publicamente na página do CELGA²³, naquela que foi a página primitiva do *Corpus de Produções Escritas de Aprendentes de PL2* até 2018. Nesta fase, as produções encontravam-se organizadas por Língua Materna, Nível de proficiência e Fase de recolha, e podiam ser consultadas e descarregadas em formato *Word* ou *text*. Os dados dos informantes podiam igualmente ser consultados e descarregados a partir de um documento em formato *Excel* contendo todas as informações recolhidas da ficha individual preenchida pelos aprendentes.

Atualmente, com a transferência dos ficheiros relativos aos dados e metadados do *corpus*, para o ambiente TEI e a sua conversão em ficheiros *XML*, as produções encontram-se disponíveis na página do CELGA-ILTEC²⁴, para consulta de livre acesso, através da plataforma TEITOK, organizados por *Nacionalidade, Língua materna, Nível de proficiência, Fase de recolha e Estímulo*, e os dados dos informantes podem ser consultados no cabeçalho de cada produção escrita. As informações relativas aos metadados são disponibilizadas nas páginas de apresentação do *corpus* e de descrição da metodologia seguida.

A partir da sua disponibilização na plataforma TEITOK, o *PEAPL2_PLE* passou a ser um *corpus* anotado com informação linguística associada. As etiquetas de natureza linguística com que as produções escritas estão anotadas são referentes à lematização e à classificação morfossintática, mas também são disponibilizadas outras etiquetas relativas às formas de representação do texto, nomeadamente, *Transcrição, Forma do aluno e Forma corrigida*.

Tendo o sistema de anotação possível no sistema TEITOK sido já apresentado no capítulo anterior, observaremos, agora, algumas das etiquetas disponibilizadas para cada produção escrita no *corpus* *PEAPL2_PLE*.

A *Transcrição* corresponde ao texto originalmente escrito pelo aprendente, sem qualquer tipo de alteração ou correção por parte do professor, respeitando as suas hesitações, correções ou adições. Tal como se pode observar na figura 20, a seleção desta opção de representação do texto permite visualizar os segmentos modificados pelo aprendente.

²³ Em www.uc.pt/fluc/rcpl2 .

²⁴ Em <http://teitok2.iltec.pt/peapl2-ple/> .

EN | PT

Home | Structure | Research | Resources | Members | Training | Activities | Contact

CELGA ILTEC
Centro de Estudos de Língüística Geral
& Aplicada da Universidade de Coimbra

PEAPL2 alemao.b1.126.50.2I

Apresentação alemao.b1.126.50.2I

Documentos

Pesquisar

Metodologia

Login

Língua materna Alemão

Gênero M

Nacionalidade Alemã

QECRL B1

mais dados

Powered by TEITOK
© Maarten Janssen,
2014.

Opções de representação

Texto: [Transcrição](#) | [Forma do aluno](#) | [Forma corrigida](#) | Mostrar: [Cores](#) | Etiquetas: [Classe morfosintática](#) | [Lema](#)

Meu país, Alemanha, é um país com muitos rostos diferentes. O nível cultural e geográfico depende, sem dúvida, da região e do ponto de vista. Eu tentarei, uma mesmo assim, uma descrição, que a qual vai ser necessariamente muito pessoal e subjetivo. Minha família mora no norte de Hesse, numa região muito rural e com poucos habitantes.

Em contraposição disso, eu estudo em Heidelberg a que é uma cidade universitária, vital e urbano com um clima muito urbano. Enquanto nós não encontramos na região da minha família nenhuns monumentos importantes na região da minha família (excepções: um castelo velho e uma cidade do século 17), é Heidelberg um lugar o qual os turistas conhecem bem. Depois de chegar em Heidelberg a primeira vez, eu pensei: "Aqui há mais Japoneses do que estudantes Mas isso não é verdade (acho que).

Heidelberg tem uma biblioteca geral velha e bellissima, muitos edifícios universitários bem legais, uma ponte velha e, quem não sabe, o castelo famoso. Heidelberg fica no norte de Baden – Vurtembergia, como Hesse uma das 16 regiões da Alemanha.

O que posso contar sobre hábitos significativos da minha cultura? Os alemães chegam (frequentemente) em ponto, eles gostam de discutir nos bares durante eles bebem cerveja. Na Alemanha há poucas greves e muitas festas (p.e. mercados do Natal com vinho quente, bem gostoso). Na região da minha família, as pessoas são mais fechadas e falam, especialmente em comparação com os habitantes de Heidelberg, pouco. Mas eles sabem fazer festas também, os senhores podem crer isso. E tudo que eu escrevi verdade? Nunca se sabe (Quem conta um conto acrescenta um ponto).

[Descarregar texto](#)

Figura 20: Opção de representação do texto referente à *Transcrição* de uma produção escrita no PEAPL2_PLE.

A *Forma do aluno* é a representação da versão final do aprendente, após as suas alterações, como se pode verificar pela comparação das figuras 20 e 21.

EN | PT

Home | Structure | Research | Resources | Members | Training | Activities | Contact

CELGA ILTEC
Centro de Estudos de Língüística Geral
& Aplicada da Universidade de Coimbra

PEAPL2 alemao.b1.126.50.2I

Apresentação alemao.b1.126.50.2I

Documentos

Pesquisar

Metodologia

Login

Língua materna Alemão

Gênero M

Nacionalidade Alemã

QECRL B1

mais dados

Powered by TEITOK
© Maarten Janssen,
2014.

Opções de representação

Texto: [Transcrição](#) | [Forma do aluno](#) | [Forma corrigida](#) | Mostrar: [Cores](#) | Etiquetas: [Classe morfosintática](#) | [Lema](#)

Meu país, Alemanha, é um país com muitos rostos diferentes. O nível cultural e geográfico depende, sem dúvida, da região e do ponto de vista. Eu tentarei, uma mesmo assim, uma descrição, a qual vai ser necessariamente muito pessoal e subjetivo. Minha família mora no norte de Hesse, numa região muito rural e com poucos habitantes.

Em contraposição disso, eu estudo em Heidelberg a que é uma cidade universitária, vital e urbano com um clima muito urbano. Enquanto nós não encontramos nenhuns monumentos importantes na região da minha família (excepções: um castelo velho e uma cidade do século 17), é Heidelberg um lugar o qual os turistas conhecem bem. Depois de chegar em Heidelberg a primeira vez, eu pensei: "Aqui há mais Japoneses do que estudantes Mas isso não é verdade (acho que).

Heidelberg tem uma biblioteca geral velha e bellissima, muitos edifícios universitários bem legais, uma ponte velha e, quem não sabe, o castelo famoso. Heidelberg fica no norte de Baden – Vurtembergia, como Hesse uma das 16 regiões da Alemanha.

O que posso contar sobre hábitos significativos da minha cultura? Os alemães chegam (frequentemente) em ponto, eles gostam de discutir nos bares durante eles bebem cerveja. Na Alemanha há poucas greves e muitas festas (p.e. mercados do Natal com vinho quente, bem gostoso). Na região da minha família, as pessoas são mais fechadas e falam, especialmente em comparação com os habitantes de Heidelberg, pouco. Mas eles sabem fazer festas também, os senhores podem crer isso. E tudo que eu escrevi verdade? Nunca se sabe (Quem conta um conto acrescenta um ponto).

[Descarregar texto](#)

Figura 21: Opção de representação do texto referente à *Forma do aluno* de uma produção escrita no PEAPL2_PLE.

Relativamente à *Correção ortográfica*, foram tidos em conta apenas desvios inequivocamente ortográficos, ou seja, todos os erros de natureza ortográfica que pudessem comportar erros de natureza morfosintática não foram corrigidos, sob pena de interferirem na leitura e interpretação dos dados. Os erros que resultam de leituras conjeturadas também não foram tidos em conta por poderem acarretar *nuances* de sentido que não seria possível validar. Ao optar pela visualização do

texto corrigido, podemos, ainda, recorrer ao código de cores (vermelho) para mais facilmente detetar as formas corrigidas, como se pode observar na figura 22:

The screenshot shows the PEAPL2 interface for a document titled 'alemao.b1.126.50.2I'. The 'Opções de representação' (Representation Options) section is active, with 'Forma corrigida' (Corrected Form) selected. This option highlights orthographic corrections in red. The text displayed is in German and discusses the cultural and geographical aspects of Germany and the region of Hesse. The interface includes navigation tabs like 'Transcrição', 'Forma do aluno', and 'Forma corrigida', and a search bar for 'Cores' and 'Etiquetas'.

Figura 22: Opção de representação do texto referente à Correção ortográfica de uma produção escrita no PEAPL2_PLE.

Relativamente à lematização, no PEAPL2_PLE deve ter-se em conta que as palavras compostas correspondem a apenas um lema; já as formas contraídas, como é o caso de preposições com determinantes ou pronomes, e a associação de verbo e pronome pessoal clítico, correspondem a dois lemas. Também os sinais de pontuação foram lematizados, à exceção do hífen. Ao seleccionar-se a opção referente à lematização, a informação disponibilizada é a que se pode observar no seguinte exemplo:

The screenshot shows the PEAPL2 interface for a document titled 'alemao.b1.126.50.2I'. The 'Opções de representação' (Representation Options) section is active, with 'Lema' (Lemma) selected. This option displays a detailed linguistic annotation of the text, showing the lemmatized form of each word and its grammatical category. The text displayed is in German and discusses the cultural and geographical aspects of Germany and the region of Hesse. The interface includes navigation tabs like 'Transcrição', 'Forma do aluno', and 'Forma corrigida', and a search bar for 'Cores' and 'Etiquetas'.

Figura 23: Anotação linguística referente à lematização de um texto no PEAPL2_PLE.

Cada palavra, entendida como unidade de significado, possui uma etiqueta que a identifica quanto à sua classe e subclasse morfofossintáticas, flexão nominal em número e valor de gênero e flexão verbal (tempo e modo). Os sinais de pontuação lematizados também estão anotados com etiquetas.

The screenshot shows the PEAPL2_PLE interface. At the top, there is a header for CELGA ILTEC (Centro de Estudos de Linguística Geral e Aplicação da Universidade de Coimbra). Below the header, there are navigation links: Home, Structure, Research, Resources, Members, Training, Activities, Contact. The main content area displays the text 'alemão.b1.126.50.21' and 'alemão.b1.126.50.21'. There are sections for 'Língua materna' (Alemão), 'Gênero' (M), 'Nacionalidade' (Alemã), and 'QECL' (B1). Below this, there are 'Opções de representação' (Transcrição, Forma do aluno, Forma corrigida) and a list of linguistic tags. A tooltip is visible over the word 'região', showing its class 'Classe morfofossintática' and name 'Nome (NCF5000) comum; feminino; singular'.

Figura 24: Anotação linguística referente à classe morfofossintática de um texto no PEAPL2_PLE.

As etiquetas de natureza linguística podem ser seleccionadas uma a uma ou combinadas entre si e, ainda, combinadas com as diferentes opções de representação textual, como se pode observar na figura 25.

The screenshot shows the PEAPL2_PLE interface. At the top, there is a header for CELGA ILTEC (Centro de Estudos de Linguística Geral e Aplicação da Universidade de Coimbra). Below the header, there are navigation links: Home, Structure, Research, Resources, Members, Training, Activities, Contact. The main content area displays the text 'alemão.b1.126.50.21' and 'alemão.b1.126.50.21'. There are sections for 'Língua materna' (Alemão), 'Gênero' (M), 'Nacionalidade' (Alemã), and 'QECL' (B1). Below this, there are 'Opções de representação' (Transcrição, Forma do aluno, Forma corrigida) and a list of linguistic tags. A tooltip is visible over the word 'região', showing its class 'Classe morfofossintática' and name 'Nome (NCF5000) comum; feminino; singular'.

Figura 25: Visualização, em simultâneo, das opções de anotação linguística de um texto no PEAPL2_PLE (*Forma corrigida, Classificação morfossintática e Lema*).

Importa referir que, num *corpus* aberto e acessível, todas as etapas da sua atualização são fundamentais na descrição dos dados que o sustentam. Por esta razão se manteve ativa a página inicial do projeto para eventual consulta dos dados e metadados relativos à fase anterior à da sua transferência para o ambiente TEI (por exemplo, para aceder às produções escritas em formato *word* ou ao perfil do aprendentes em ficheiro *excel*).

3.2. *Corpus* Oral de Português L2 (COraI-Co)

O *Corpus* Oral de PL2 (COraI-Co)²⁵ é um projeto do CELGA-ILTEC, desenvolvido na Faculdade de Letras da Universidade de Coimbra. Trata-se de um *corpus* de produções orais de aprendentes de português língua não materna, recolhido nos anos de 2014 e 2015, cujo principal objetivo é constituir um acervo que sustente empiricamente, não só a investigação realizada no âmbito do processo de ensino-aprendizagem de PL2, com relevância na construção da(s) interlíngua(s) dos aprendentes, mas também opções metodológicas e a construção de materiais didáticos.

O *corpus* disponibiliza ficheiros áudio referentes às produções de 56 informantes, distribuídos por 24 línguas maternas e por níveis de proficiência entre o A1 e o C1(+), a frequentarem cursos ou unidades curriculares de ensino do português para estrangeiros na Faculdade de Letras de Coimbra e que, por isso, combinam experiências de aprendizagem em contexto formal e em situações de imersão. As informações relativas aos informantes foram obtidas através de questionário sobre dados pessoais e sobre o percurso de aprendizagem da língua materna e de outras línguas não maternas por parte dos aprendentes²⁶.

Cada um dos informantes produziu textos orais de natureza diversa em função de diferentes estímulos agrupados em dois domínios: produção oral e leitura oral. Assim, as produções orais foram obtidas a partir da realização das seguintes tarefas: entrevista semiestruturada (tarefa 1), elicitación de atos ilocutórios - pedido, convite/sugestão, censura, agradecimento, pedido de desculpas, elogio/felicitações - (tarefa 2), construção de um texto narrativo a partir de uma sequência de imagens (tarefa 3) e nomeação de figuras a partir de um suporte pictórico (tarefa 4). Os dados resultantes da leitura oral foram obtidos a partir da leitura de um texto (tarefa 5) e de duas listas de palavras (tarefas 6 e 7). No que concerne à tarefa 2, elicitación de atos ilocutórios, aos alunos do nível A1 não foram apresentadas situações relativas aos atos ilocutórios de *censura* nem de *elogio/felicitações*. Também na tarefa 3, e para este mesmo grupo de alunos, se abreviou a sequência de imagens. É possível verificar que, para alguns informantes, nem todas as tarefas estão disponíveis para consulta porque, embora as tenham realizado, considerou-se que o ficheiro áudio não tinha a qualidade pretendida em virtude de fatores externos à gravação (ruídos exteriores,

²⁵ O COraI-Co pode ser consultado em <http://teitok2.iltec.pt/coralco/index.php?action=home>.

²⁶ Ao questionário usado no PEAPL2, acrescentou-se uma questão destinada à identificação das variedades do português com as quais os aprendente entraram em contacto ao longo do seu percurso de aprendizagem.

discurso inteligível). Todas estas informações relativas ao protocolo de recolha de dados encontram-se disponíveis para consulta nas páginas de *Descrição* do COral-Co²⁷.

Na fase de preparação do *corpus*, tomaram-se algumas opções metodológicas relacionadas com o processo de transcrição em si e com as especificidades do registo oral.

À exceção das tarefas de leitura (tarefas 5, 6 e 7), que, pela sua natureza, estão associadas a texto previamente redigido, as tarefas de produção oral, que se encontram em fase de transcrição, serão todas transcritas observando a ortografia convencional do português europeu, não havendo lugar à transcrição fonética. No entanto, recorreu-se a formas gráficas desviantes para assinalar casos em que a estrutura fonológica da palavra é percecionada, com clareza, pelo transcritor como não sendo coincidente com a língua-alvo. Nestes casos, a palavra é transcrita tal como é percecionada, mas dando origem a duas situações distintas: formas que foram identificadas para assinalar quer o recurso a palavras quer a ocorrência de pequenos excertos em outra língua, que não a língua-alvo, etiquetados como estrangeirismos (cf. figura 26) e formas com uma estrutura que se aproxima de formas existentes na língua-alvo (cf. figura 27). Neste último caso, as formas não foram corrigidas ou assinaladas, pois a transcrição de alguns desvios permite observar aspetos da aquisição fonológica e a disponibilização do áudio permite ao utilizador a confirmação.

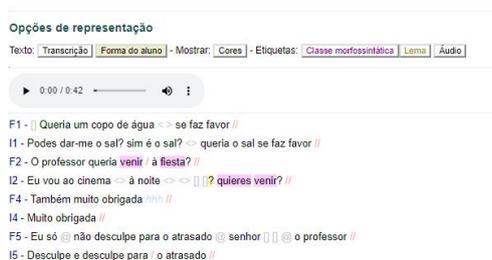


Figura 26: Exemplos de formas gráficas desviantes da língua-alvo, identificadas como estrangeirismos, na transcrição de uma produção oral do COral-Co.

²⁷ Consultar em <http://teitok2.iltec.pt/coralco/index.php?action=estimulos>.

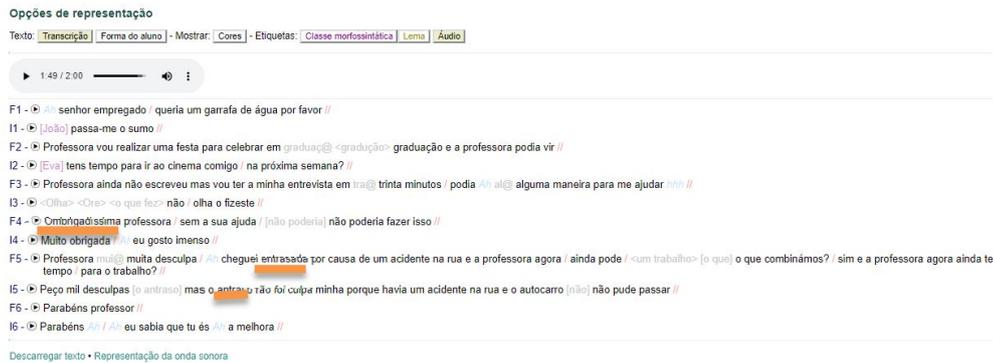


Figura 27: Exemplos de formas gráficas desviantes da língua-alvo na transcrição de uma produção oral do COral-Co.

Em situações em que não é perceptível, com clareza, a forma produzida pelo informante, esta é assinalada na transcrição como uma interpretação conjecturada e, nos casos em que é impossível percecionar a forma produzida, esta assinala-se como sendo ininteligível.

Foram igualmente transcritos todos os segmentos paralinguísticos característicos do registo oral, como hesitações, repetições, reformulações, interrupções, bem como foram assinalados todos os outros elementos de natureza extralinguística (risos, sons) que, por vezes, espontaneamente, ajudam a estruturar o enunciado. Também se mantiveram, devidamente assinaladas, nas tarefas em que tal era relevante, as intervenções do entrevistador e assegurou-se o anonimato dos informantes²⁸, ocultando qualquer informação que pudesse conduzir à sua identificação.

Ainda relativamente ao processo de transcrição, há a referir que os únicos aspetos de natureza prosódica transcritos foram a interrogação (pelo recurso ao ponto de interrogação), dada a relevância desse aspeto prosódico em estudos de natureza discursivo-pragmática e as pausas (breve e longa) que foram assinaladas com recurso a convenções de outro tipo.

Na transcrição das produções orais, respeitaram-se, então, as convenções que a seguir se sistematizam.

²⁸ Este procedimento também se manteve nos ficheiros áudio, tendo sido adotado um sinal sonoro para o efeito.

Convenções adotadas na transcrição de produções orais	
/	Pausa curta
//	Pausa longa
<i>Uhm</i> <i>Ah</i> <i>Eh</i> <i>Oh</i>	Pausas preenchidas
hhh	Risos
xxx@	Truncamento
[xxx]	Repetição
<xxx>	Reformulação
[...]	Palavra ou segmento ininteligível
xxx	Estrangeirismo
xxx	Pouco claro. Existência de dúvida
<i>Eva</i> <i>João</i>	Nome genérico para mulheres Nome genérico para homens (Ocultação de elementos passíveis de reconstituir a identidade do falante)
INT (Tarefa 3) :XXX (Tarefa 2)	Segmentos onde se pode ouvir a voz do entrevistador
<i>Bater de dedos</i>	Cinesia

Tabela 2: Convenções adotadas, no COral-Co, na transcrição de produções orais.

As produções orais foram transcritas no ambiente TEITOK, possibilitando o alinhamento entre os segmentos áudio e a respetiva transcrição. Assim, ao aceder às produções, é possível ouvir integralmente o ficheiro que corresponde à realização da tarefa, ou, opcionalmente, ouvir apenas segmentos previamente delimitados (cf. figura 28). A audição dos ficheiros pode ser também acompanhada pela visualização da onda sonora (cf. figura 29). Qualquer que seja a opção, os ficheiros áudio encontram-se sempre seguidos das transcrições, na mesma página.



Figura 28: Opções de audição do ficheiro áudio que acompanha transcrição, no COraL-Co.



Figura 29: Representação da onda sonora, seguida da transcrição da produção, no COraL-Co.

O COraL-Co é um *corpus* anotado que permite analisar as produções em função de dois níveis de anotação considerados relevantes para este *corpus*, nomeadamente, ao nível do texto - i. *Transcrição*, ii. *Forma do aluno* e iii. *Áudio* - e a nível linguístico - iv. *Classe morfosintática* e v. *Lema*.

i. *Transcrição*

Ao seleccionar a opção *Transcrição*, o utilizador tem acesso à transcrição da produção original do falante, com todas as suas hesitações, reformulações, truncamentos e segmentos característicos do discurso oral, tal como referido anteriormente.

ii. *Forma do aluno*

Já a *Forma do aluno* permite observar a transcrição “limpa” de todos estes segmentos, como se se tratasse da versão “alvo” do aluno. Note-se que o ficheiro áudio corresponde à transcrição e não à

forma do aluno, uma vez que a produção oral é a original a partir da qual se faz a transcrição, não havendo manipulação do ficheiro áudio a este nível.

iii. Áudio

Ao selecionarmos a opção *Áudio*, tal como já foi referido, ativamos a possibilidade de escolher quais os segmentos que queremos ouvir, isolados da produção integral. Nesta situação, ao ouvir os segmentos um a um, por vezes, é possível ouvir, no final de um segmento, o início do seguinte. Esta foi uma opção ponderada aquando da segmentação dos trechos áudio, pois como é próprio do discurso oral, as unidades nem sempre apresentam pausas suficientemente longas na sua delimitação alinhando-se num *continuum* fonético. Nestes casos, considerou-se preferível ouvir um pouco além do pretendido do que não ouvir na sua totalidade o segmento em análise.

Relativamente às anotações de natureza linguística, recordamos que os textos foram previamente tokenizados e lematizados e só depois as unidades foram classificadas de acordo com a categoria morfossintática a que pertencem. No processo de lematização, deve ter-se em conta que não foram lematizadas as unidades que se referem a: a) segmentos sonoros típicos do discurso oral (correspondendo a hesitações, reformulações, truncamentos, risos, pausas), nem a b) elementos prosódicos como as pausas não preenchidas (breve e longa).

Opções de representação

Texto: Transcrição Forma do aluno - Mostrar: Cores - Etiquetas: Classe morfossintática Lema Áudio

0:00 / 3:10

INF - A mulher estava a andar na rua e ela encontrou o roubo //

INF - E o roubo //

INF - O roubo trazia a e é mala ? //

INT - Sim //

INF - A mala de A mala dela e //

INF - Mas o roubo não sabia que a polícia é //

INF - A pai da O pai dela era polícia e //

INF - Quando o roubo //

INF - [...] //

INF - Turned his back ? Turned his face ? //

Figura 30: Exemplo de anotação linguística (classe morfossintática e lema) de uma produção no COraL-Co.

As informações que constituem a documentação do *corpus* - objetivos, informantes, estímulos e convenções de transcrição - encontram-se em secções distintas do *corpus* às quais é possível aceder através do menu lateral de navegação da página.

O *corpus* está disponível para pesquisa através da plataforma TEITOK, permitindo fazer pesquisas a dois níveis. A nível do *corpus*, com base nas anotações textual e linguística, podemos pesquisar dados referentes aos textos que resultam da transcrição, e, simultaneamente, consultar os ficheiros áudio que lhes deram origem. A nível da estrutura que sustenta o *corpus*, é possível fazer pesquisas por *informante*, *tarefa*, *proficiência*, *nacionalidade*, *língua materna* e, no que respeita especificamente à tarefa 2, pode pesquisar-se por *subtarefa* (combinando ato ilocutório e situação), por *situação (formal/informal)* e por *ato ilocutório*. Para além do preenchimento destes campos, também existe a possibilidade de realizar uma pesquisa no *corpus* por formas que foram alvo de processos de transferência da L1 dos aprendentes - *Foreign Text*. Observe-se, a este propósito, a área de pesquisa do COral-Co disponibilizada através da plataforma TEITOK (figura 31).

Figura 31: Área de pesquisa do COral-Co.

Através da plataforma TEITOK, é possível também descarregar os resultados de pesquisa e as frequências das ocorrências, em formato *excel*, os textos transcritos, em formato *text*, e os ficheiros áudio, em formato *mp3*.

PARTE II - O PROJETO

INTRODUÇÃO

A segunda parte deste projeto pretende dar a conhecer o trabalho de natureza prática, sustentado pelo enquadramento realizado na primeira parte, que permitiu desenvolver um conjunto de FAQ e respetivas respostas, para a exploração de valências dos *corpora* informatizados de aprendentes PEAPL2_PLE e COral-Co.

No primeiro capítulo, será apresentado o projeto naquilo que são as suas linhas orientadoras para a definição das FAQ acima enunciadas e, complementarmente, será descrita a metodologia referente às várias etapas que constituíram este trabalho, bem como os procedimentos adotados nestas etapas.

O segundo capítulo será inteiramente dedicado às FAQ, desde a descrição dos fundamentos que presidiram à sua elaboração, passando pela sua apresentação, e culminando com os procedimentos observados na construção de respostas e na disponibilização dos materiais construídos na plataforma TEITOK.

Quanto à elaboração das FAQ, teve-se em consideração as áreas de investigação no âmbito do português língua não materna, para a definição de exemplos de eventuais questões de pesquisa baseadas em *corpora* de aprendentes. Para além disso, a elaboração das FAQ incidiu sobre três domínios que são fundamentais para a extração consistente de dados fiáveis de um *corpus*: a anotação a que foi submetido, as pesquisas que permite realizar e o armazenamento dos dados e metadados que disponibiliza.

As respostas às FAQ pautaram-se por uma linguagem clara e objetiva, rigorosa contudo, a pensar nas orientações de pesquisa e nos prováveis utilizadores do PEAPL2_PLE e do Coral-Co. Para tornar a pesquisa mais intuitiva para os seus utilizadores, permitindo explorar todas as valências dos *corpora* informatizados de aprendentes, as respostas foram construídas em dois formatos distintos: *PDF* e *mp4*. Nesta secção, daremos, também, conta dos procedimentos adotados na disponibilização destes ficheiros nas páginas de pesquisa *online* de cada um dos *corpora*.

CAPÍTULO 1 - APRESENTAÇÃO DO PROJETO E METODOLOGIA

INTRODUÇÃO

O primeiro capítulo tem por objetivos principais apresentar o projeto desenvolvido no âmbito da exploração de valências em pesquisa de *corpora* informatizados de produções orais e escritas de aprendentes e descrever a metodologia utilizada para a sua concretização.

Quanto ao projeto, serão apresentadas as linhas orientadoras que estão na sua génese - a natureza das pesquisas e o perfil dos utilizadores dos *corpora*, a organização e disponibilização dos dados e metadados e a exploração de valências da plataforma *online* para pesquisa de dados empíricos - e os objetivos específicos inerentes à conceção dos materiais a disponibilizar na plataforma TEITOK - tornar a pesquisa mais intuitiva, facilitar o acesso aos dados e metadados e explorar as opções de armazenamento de dados.

As etapas serão descritas em função dos diferentes momentos de desenvolvimento do projeto e dos procedimentos adotados em cada uma delas para a definição das *Frequently Asked Questions* (FAQ) e para a construção das respetivas respostas. Assim, descrevem-se seis etapas: pesquisa de áreas de investigação no âmbito do português língua não materna, testagem de eventuais perguntas de investigação na plataforma TEITOK, otimização da plataforma, organização e elaboração das FAQ por domínios de pesquisa, em função das dificuldades (ainda) sentidas na pesquisa de *corpora* na plataforma *online*, construção das respostas às FAQ em suportes distintos (*PDF* e *mp4*) e, finalmente, disponibilização dos materiais construídos (texto informativo de introdução à pesquisa e conjunto de FAQ e respetivas respostas) nas páginas de pesquisa dos respetivos *corpora*.

1.1. Apresentação do projeto

Este projeto assenta na multiplicidade de valências que a plataforma TEITOK oferece ao nível da exploração de dois *corpora* disponibilizados pela Faculdade de Letras da Universidade de Coimbra no âmbito de dois projetos afins do CELGA-ILTEC: o COral-Co (*Corpus Oral de Português L2 - Coimbra*) e o subcorpus de *Português Língua Estrangeira (PLE)*, que integra o PEAPL2 (*Corpus de Produções Escritas de Aprendentes de L2*).

Embora os dois *corpora* sejam de natureza distinta, ambos reconhecem a «importância da sustentação empírica no âmbito da investigação sobre aprendizagem/aquisição de língua não materna e sobre a construção e definição das interlínguas dos aprendentes que têm o português como língua-alvo.»²⁹ Assim, a reflexão em torno do pressuposto de que os *corpora* são instrumentos de trabalho que facultam dados empíricos para a descrição de determinadas estruturas da língua, bem documentados e pesquisáveis, aplicada aos *corpora* de aprendentes de português língua não materna, foi o ponto de partida para estabelecer os objetivos para o trabalho a desenvolver:

- i. Considerando os objetivos subjacentes à criação dos dois *corpora*, definir os possíveis utilizadores e as eventuais questões de pesquisa que os dados poderão ajudar a esclarecer;
- ii. Dando conta do trabalho prévio de recolha e preparação das produções que constituem o *corpus*, definir e organizar a informação a disponibilizar ao nível dos textos em si e ao nível dos metadados;
- iii. Explorar a plataforma e orientar o seu uso para pesquisa e, consequentemente, promover a máxima rentabilização na extração de informação fiável dos *corpora*.

Estas linhas orientadoras conduziram à definição de um conjunto de questões cujas respostas visam, de uma forma genérica, aproximar o utilizador da plataforma ao ambiente de pesquisa e, por conseguinte, ao *corpus* que analisa. Assim, no âmbito deste projeto, a conceção dos materiais resultantes de um trabalho prévio, e exaustivo, de testagem de ferramentas pretende a) tornar a pesquisa mais intuitiva, por um lado, simplificando procedimentos, mas, simultaneamente fornecer ao utilizador estratégias de pesquisa mais avançadas sempre que a investigação o exija; b) facilitar o acesso aos metadados dos *corpora*; c) explorar as possibilidades de armazenamento dos dados, em diversos formatos. Assim, será possível tirar o máximo partido de todas as potencialidades da plataforma.

²⁹ Apresentação do COral-Co, (coord. Isabel Santos), em <http://teitok2.iltec.pt/coralco/index.php?action=home>.

Elaborou-se, então, para cada um dos *corpora*, um conjunto de questões, de que daremos conta no capítulo seguinte, cujas respostas pretendem orientar os utilizadores nas suas pesquisas.

1.2. Metodologia

Após a sistematização dos objetivos inerentes à elaboração de questões, o projeto foi desenvolvido nas diferentes etapas que a seguir se descrevem.

I

Tendo presente os objetivos inerentes à criação do PEAPL2_PLE e do COral-Co, já referidos ao longo deste trabalho, realizou-se, numa fase inicial, uma recolha de temas de investigação que pudessem recorrer à análise dos dados destes dois *corpora*. Como tal, foi consultada a “*Bibliografia sobre aquisição, aprendizagem e ensino do Português Europeu como Língua não Materna*” (org. Cristina Martins) da Cátedra de Português Língua Segunda Língua Estrangeira (Instituto Camões/Universidade Eduardo Mondlane)³⁰. Com este mesmo objetivo foram, ainda, reunidos títulos referentes aos anos de 2018 e 2019 que viriam a integrar essa mesma bibliografia, contribuindo para a sua atualização.

Desta etapa de trabalho resultou uma recolha de temas de investigação relevantes na área da aquisição, aprendizagem e ensino do português língua não materna, a partir dos quais foi possível equacionar questões de pesquisa e começar a testar a recolha de dados relevantes na plataforma TEITOK, dando, assim, início à segunda etapa do projeto. Estes temas de investigação serão apresentados no capítulo 2 (secção 2.1. A.).

II

A primeira abordagem à plataforma TEITOK, na perspetiva do utilizador, consistiu em inserir expressões de busca simples, preencher/selecionar campos pré-definidos e observar os resultados obtidos. O principal objetivo foi testar todas as funcionalidades de que a plataforma dispõe e verificar quais as pesquisas que permite, ou não, realizar. Após aferição dos resultados dos testes, procedeu-se à listagem de todas as situações passíveis de causar dúvidas e de conduzir a resultados menos consistentes, ou insuficientes, como resposta a uma dada pesquisa, entre as quais: opções inoperacionais e outras pouco intuitivas, inexistência de campos referentes à pesquisa de alguns

³⁰ Esta bibliografia pode ser consultada em <https://www.catedraportugues.uem.mz/?target=lista-bibliografia-aquisicao>.

dados/metadados existentes nos *corpora*, dificuldades no processo de importação de dados e área de pesquisa pouco intuitiva.

III

Face a algumas inconsistências detetadas, foi necessário redefinir a metodologia do trabalho e introduzir uma nova etapa de otimização da plataforma.

Tal como havíamos referido na primeira parte deste trabalho, aquando da descrição das valências do TEITOK (Parte I, cap. 2), a pesquisa, embora esteja direcionada para o utilizador externo, é concebida pelo administrador da plataforma e as suas funcionalidades são atualizadas em função da informação nela inserida, pelo utilizador interno, aquando da preparação do *corpus* para posterior disponibilização. Saliéntamos, ainda, que a pesquisa é o reflexo deste trabalho prévio de preparação dos dados e dos metadados. Então, se a informação está armazenada no *corpus*, é expectável que ela possa ser pesquisável.

Na nova etapa de otimização da plataforma, as áreas de intervenção de que esta foi alvo, e que serão descritas no capítulo 2 (secção 2.1. B), abrangeram dois níveis distintos: o nível do *interface*, na ótica do utilizador; e o nível estrutural, na ótica do administrador. Alguns destes aspetos prendem-se, como se compreende, com a recente disponibilização dos *corpora* na plataforma TEITOK e a necessária atualização do CQP para pesquisa.

Por sua vez, esta etapa foi fundamental na redefinição de questões que pudessem ser relevantes no auxílio à pesquisa de *corpora*.

IV

Após o trabalho de testagem de pesquisas baseadas em questões de investigação relevantes no âmbito da análise de *corpus* e o reconhecimento e otimização da plataforma, foi possível listar os casos que i) poderiam suscitar dúvidas, ii) necessitariam de esclarecimentos adicionais/complementares, iii) beneficiariam de um escopo de pesquisa mais avançado, iv) careciam de exemplificação ou v) exigiam orientação prévia.

Realizada esta triagem, constatámos que, para dar resposta a estas demandas, seria necessário elaborar questões de orientação de pesquisa organizadas em três níveis distintos: I. anotação do *corpus*, II. construção de pesquisa e III. armazenamento de dados. Dentro de cada uma destas

secções, as perguntas encontram-se ordenadas e numeradas sugerindo um possível percurso de pesquisa.

Deu-se, então, início à redação das questões, doravante FAQ (*Frequently Asked Question/s*). Como se pode constatar pelo que fica descrito na metodologia, estas questões não resultam de uma recolha, efetiva, de questões por parte de utilizadores da plataforma, uma vez que essa funcionalidade não está, para já, disponível. Resultam, sim, das dificuldades encontradas na exploração das valências de pesquisa da plataforma TEITOK que levámos a cabo, sempre na perspetiva do utilizador externo, e que, por isso, poderão constituir um banco de questões úteis, não só para o investigador especializado, mas também para o utilizador comum. As FAQ serão apresentadas na secção 2.2. do capítulo 2 desta parte II.

V

As respostas às FAQ foram construídas em dois formatos: *PDF* e *mp4*. Esta opção ficou a dever-se a questões de ordem técnica, por um lado, mas também, e sobretudo, a pensar na ótica do utilizador externo. Estes materiais foram concebidos para serem um auxiliar na exploração dos *corpora* e, como tal, pretende-se que disponibilizem a informação de forma clara, objetiva e sistematizada.

Os ficheiros *PDF* têm a vantagem de serem sucintos e facilmente armazenados pelo utilizador para uma consulta rápida, a qualquer momento. A descrição da pesquisa é feita *passo a passo* e ilustrada com imagens da construção de pesquisa e/ou dos resultados obtidos. Em alguns casos, são disponibilizadas pequenas notas, para auxiliar o entendimento das respostas dadas, e existem remissões, através de *links*, para outros documentos, para outras secções da página ou para outras páginas dos *corpora*.

Os ficheiros em formato *mp4* são uma mais-valia na medida em que permitem *ver fazer*, esclarecendo dúvidas que possam surgir da leitura da versão escrita dos documentos em *PDF*. Estes vídeos explicativos oferecem a possibilidade de visualizar todos os passos da pesquisa, em tempo real, através da gravação do écran do computador, acompanhados, em simultâneo, da descrição áudio das ações desenvolvidas.

O nome atribuído aos ficheiros, quer em formato *PDF* quer em formato *mp4*, obedeceu à seguinte estrutura: **FAQ número da questão_tópico da questão**, para que possam ser mais facilmente identificados. Por exemplo: **FAQ1_anotação do corpus**.

VI

Na fase final do trabalho, foi elaborado e disponibilizado, na página de pesquisa dos respetivos *corpora*, um texto de introdução à pesquisa e às FAQ, com informações genéricas a ter em conta aquando de pesquisas levadas a cabo na plataforma TEITOK.

Por fim, após ter sido testada a compatibilidade dos materiais produzidos com a plataforma e ter sido verificado o seu correto funcionamento, procedeu-se ao *upload* dos ficheiros *PDF* e *mp4* que se encontram disponíveis para consulta, sem restrições de acesso, nas páginas de pesquisa do PEAPL2_PLE e do COral-Co.

CAPÍTULO 2 - RESULTADOS - FAQ E RESPOSTAS

INTRODUÇÃO

O segundo capítulo visa, num primeiro momento, a apresentação das FAQ em função dos fundamentos que presidiram à sua elaboração: as áreas de investigação no âmbito do português língua não materna e os domínios de pesquisa - anotação do *corpus*, pesquisa e armazenamento de dados. Assim, será apresentado um conjunto de FAQ para cada um dos *corpora*, PEAPL2_PLE e COraL-Co, e um quadro-síntese com as FAQ distribuídas por áreas de investigação e por domínios de pesquisa.

Num segundo momento, terá lugar a descrição dos procedimentos levados a cabo na construção das respostas às FAQ e na sua disponibilização na plataforma *online* TEITOK. As respostas foram elaboradas em dois formatos distintos (*PDF* e *mp4*), mas complementares, com vista à consulta de orientações de pesquisa em diferentes domínios, numa linguagem clara e objetiva, de forma a esclarecer as possíveis dúvidas dos utilizadores da plataforma.

Após a definição das FAQ e a construção das respostas, os materiais resultantes destas etapas foram disponibilizados nas páginas de pesquisa de cada um dos *corpora*, na plataforma TEITOK, tendo sido redigido um texto informativo que serve de introdução à pesquisa e que, por essa razão, antecede a consulta dos materiais disponibilizados.

2.1. Fundamentos para a elaboração das FAQ

O conceito de *corpus* informatizado e anotado que apresentámos na primeira parte deste trabalho conduziu à definição das linhas orientadoras para a formulação de um conjunto de FAQ destinadas a orientar a pesquisa no PEAPL2_PLE e no COral-Co, com o objetivo de extrair dados válidos para a investigação.

Assim, e de acordo com a metodologia apresentada no primeiro capítulo, começaremos por fundamentar a elaboração das FAQ em função de **A.** áreas de investigação no âmbito do português língua não materna e **B.** domínios de pesquisa, que apresentámos na etapa IV da metodologia, nomeadamente I. anotação do *corpus*, II. construção da pesquisa e III. armazenamento de dados.

A. Áreas de investigação no âmbito do PLNM

A consulta e recolha de bibliografia especializada na área da aquisição, aprendizagem e ensino do português língua não materna, tal como se descreveu na metodologia, conduziu à elaboração de uma listagem de temas de investigação que a seguir se apresenta.

Áreas de investigação no âmbito do português língua não materna
✧ O perfil do aprendente de PL2
✧ A interferência/transferência da LM na aprendizagem do PL2
✧ Estádios da interlíngua
✧ A importância do erro
✧ Exposição à língua-alvo
✧ A importância do <i>input/output</i>
✧ Áreas críticas da língua-alvo em diferentes domínios (morfologia, sintaxe, fonética, semântica, pragmática)
✧ Padrões de aquisição
✧ Estudos de análise contrastiva
✧ Os usos e os contextos
✧ A aquisição da leitura
✧ A aquisição do léxico
✧ A expressão escrita na aula de PLE

-
- | |
|---|
| <ul style="list-style-type: none">✧ A Entrevista oral inicial como instrumento de trabalho em PLNM✧ A competência oral: uma abordagem por tarefas✧ Construção de materiais didáticos✧ O QECRL✧ Múltiplos olhares sobre Portugal, o português e os portugueses✧ O ensino-aprendizagem do PLE noutros países |
|---|

Tabela 3: Áreas de investigação no âmbito do PLNM (com base na consulta da “*Bibliografia sobre aquisição, aprendizagem e ensino do Português Europeu como Língua não Materna*”).

Como se pode observar, o aprendente, a língua-alvo e o contexto de ensino-aprendizagem constituem os principais focos da investigação no domínio do processo de aquisição/aprendizagem do português língua não materna. Este foi o ponto de partida para o esboço de algumas FAQ que pudessem ajudar à pesquisa de dados relacionados com alguns destes temas de investigação, como se poderá constatar mais adiante.

No entanto, e como vimos na primeira parte deste trabalho (Parte I, cap. 1), os dados resultantes da pesquisa de *corpora* são válidos em diversas áreas e, como tal, não é possível, ao longo da vida de um *corpus*, prever quantos utilizadores, e para que fins, a eles irão aceder - investigação, didática, construção de materiais, otimização de *software*. Assim, e uma vez que o PEAPL2_PLE e o COral-Co pretendem ser acervos universalmente acessíveis, na elaboração das FAQ tivemos o cuidado de orientar os exemplos de pesquisa em função das áreas de investigação atrás elencadas, mas não apresentando casos demasiadamente complexos ao ponto de impedir o utilizador, não especialista em Linguística, de as compreender cabalmente e de tirar partido das orientações.

Neste sentido, o principal objetivo das FAQ, orientadas para questões de investigação, mas, simultaneamente, acessíveis, do ponto de vista da linguagem utilizada e dos exemplos apresentados, é fornecer as ferramentas para que seja possível tirar o máximo partido do que os dois *corpora* podem oferecer. Este mesmo cuidado foi tido em conta na forma como as respostas foram elaboradas, numa linguagem clara, simples, ainda que rigorosa e objetiva.

B. Domínios de pesquisa

I. Anotação do *corpus*

«[U]m bom corpus, do ponto de vista de quem o constrói e do ponto de vista de quem o utiliza, é um *corpus* bem documentado» (Freitas, 2015:34) porque da informação que o *corpus* disponibiliza depende a interpretação dos dados. É imperativo, por isso, que a informação seja de simples acesso e facilmente identificável.

Um dos aspetos a ter em conta no que concerne à documentação do *corpus* é a anotação. Tratando-se, nos dois casos, de *corpora* anotados, é imprescindível descrever os diferentes tipos de anotação de que foram alvo para orientar a pesquisa em função desta informação. Não é possível recolher dados consistentes através de uma pesquisa por lema se não se souber, por exemplo, o que foi lematizado; da mesma forma, não é possível fazer uma correta leitura das produções desconhecendo as convenções de transcrição. Também aqui se optou por elaborar uma questão cuja resposta permitisse descrever a anotação a que foi submetido cada um dos *corpora*, uma vez que esta informação ainda não se encontrava disponível.

Quer no PEAPL2_PLE quer no COral-Co, as informações disponibilizadas encontram-se organizadas a dois níveis e, por isso mesmo, em locais diferentes do *corpus*. A informação de carácter mais genérico de cada um dos *corpora* encontra-se disponível nas respetivas páginas de apresentação e inclui informação acerca da sua dimensão e da metodologia utilizada na recolha de produções orais e escritas. Já os dados referentes ao perfil dos aprendentes encontram-se nos cabeçalhos das produções. No entanto, o COral-Co disponibiliza esta informação detalhada em secção própria e o PEAPL2_PLE possibilita o acesso aos perfis de todos os aprendentes, organizados num ficheiro alojado na página primitiva do projeto de *Recolha de Produções de PL2*. Estando, assim, muita desta informação dispersa, houve necessidade de criar uma FAQ cuja resposta facilitasse o acesso aos metadados, quer através de remissões quer através de sistematização.

II. Construção da pesquisa

À luz da era digital que é o século XXI, é fundamental saber manipular os meios tecnológicos à disposição para a exploração de *corpora*. O facto de, com um clique, podermos extrair informação de um *corpus* de grandes dimensões é bastante atrativo, mas exige os meios e os meios devem ser constantemente atualizados e testados.

Ao realizarmos os primeiros testes na plataforma TEITOK, através do CQP, pudemos constatar que algumas funcionalidades de pesquisa ainda não se encontravam no ponto de serem utilizadas corretamente. Quanto a nós, isso ficou a dever-se a dois fatores de natureza distinta: a recente disponibilização destes *corpora* informatizados através do TEITOK e, subsequentemente, a necessidade de (re)ajustar o sistema de pesquisa à realidade de cada um.

Por este motivo, tal como se anunciou no capítulo relativo à apresentação da metodologia seguida neste projeto, houve necessidade de levar a cabo um trabalho de otimização da plataforma, a nível do sistema, isto é, na ótica do administrador, com vista aos objetivos de pesquisa delineados para cada *corpus*. Assim, as tarefas realizadas neste âmbito, e que a seguir listamos, só foram possíveis graças à estreita colaboração com Maarten Janssen, que detém a autoria da plataforma TEITOK e que é responsável pela sua manutenção e pela definição das autorizações de acesso.

1. Tradução para português da página de pesquisa (comandos, campos de pesquisa), dos cabeçalhos e do sistema anotação linguística;
2. Registo da equivalência das convenções de transcrição utilizadas no ambiente TEI;
3. Revisão das categorias morfossintáticas (eliminação de categorias irrelevantes para o português europeu);
4. Proposta de novos campos de pesquisa (“Construção de etiquetas morfossintáticas”, “Estímulo”, “Estrangeirismo”, “Tarefa”, “Ato ilocutório” e “Situação”);
5. Desbloqueio de opções de pesquisa (“Opções de frequência”, “Descarregar dados de pesquisa”, “Comparação de expressões de pesquisa guardadas”);
6. Testagem em diferentes sistemas operativos e com diferentes *browsers* para encontrar soluções comuns (descarregar dados, por exemplo);
7. Detecção de falhas de funcionamento em vários setores de pesquisa.

Relativamente às tarefas desenvolvidas para melhorar a qualidade da pesquisa disponibilizada pelo TEITOK, as 4 e 5, referentes a campos e opções de pesquisa, destacam-se pela sua relevância no que diz respeito à obtenção de dados consistentes com a informação armazenada no *corpus* e a sua preparação.

Na tarefa 4, os campos de pesquisa - “Construção de etiquetas morfossintáticas”, “Estímulo”, “Estrangeirismo”, “Tarefa”, “Ato ilocutório” e “Situação”³¹ - foram propostos em função dos dados que constituem os *corpora* e que estão disponíveis para pesquisa, mas que não eram visados nessa área. Quanto à “Construção de etiquetas morfossintáticas”, a pesquisa por categorias morfossintáticas só era possível mediante o preenchimento de um campo com expressões de pesquisa inerentes ao sistema de pesquisa do CQP, como exemplificámos na primeira parte do trabalho (parte I, cap. 2, secção 2.3.), pressupondo o conhecimento da linguagem específica de pesquisa e, por isso, muito pouco intuitiva.

Relativamente à tarefa 5, estas opções de pesquisa estão relacionadas com o armazenamento de dados e, por isso, a sua importância é crucial para a importação de dados dos *corpora*. As opções existiam na área de pesquisa, mas não estavam funcionais por razões de natureza técnica e, por esse motivo, foi necessário realizar testes para averiguar a origem do problema e chegar à solução.

Todas estas tarefas levadas a cabo com o objetivo de otimizar a plataforma permitiram corroborar o que havíamos anteriormente afirmado (Parte I, cap. 2): partindo de uma base comum - informação armazenada em ficheiros *XML* e sistema de pesquisa flexível - é possível ter uma ferramenta de pesquisa à medida de cada *corpus*.

Porém, este foi um trabalho moroso e que exigiu um conhecimento aprofundado do ambiente TEI e da plataforma TEITOK, mas que permitiu observar todas as valências da plataforma aplicadas aos *corpora* de aprendentes e averiguar quais aquelas que poderiam colocar alguma espécie de constrangimento no momento da pesquisa.

Assim, após a realização destas tarefas, pudemos verificar que a pesquisa através do *CQP* nem sempre é tão intuitiva quanto o sistema permite, na medida em há alguns aspetos, como por exemplo, a funcionalidade e a combinação de opções e até formas de pesquisa alternativas ao simples preenchimento dos campos, que carecem de alguns esclarecimentos prévios para serem utilizados corretamente e devolverem resultados consistentes. A elaboração de um conjunto de FAQ que remete diretamente para o funcionamento da construção de pesquisa pretendeu dar resposta a estes tipo de casos.

³¹ Os campos “Construção de etiquetas morfossintáticas” e “Estímulo” foram criados para responder às necessidades de pesquisa do PEAPL2_PLE, ao passo que os campos “Estrangeirismo”, “Tarefa”, “Ato ilocutório” e “Situação” se referem, especificamente, à área de pesquisa do CORal-Co.

III. Armazenamento de dados.

As FAQ orientadas para o armazenamento de dados relacionam-se com a necessidade de poder não só consultar os dados e os metadados, mas também importá-los da plataforma para facilitar o seu posterior tratamento³².

Considerando, assim, o armazenamento de dados uma etapa importante na investigação, as FAQ neste âmbito permitem esclarecer o utilizador relativamente às diversas formas possíveis de aceder aos dados e de os armazenar em ficheiros de diversos formatos. Embora muitas vezes esse procedimento não seja complexo, a verdade é que pode ser condicionado pelos diferentes *software* e motores de pesquisa *online* que o utilizador usa no seu ambiente pessoal. Por essa razão, julgámos pertinente alertar para estas situações que, por vezes, podem apresentar-se como um constrangimento.

³² A plataforma poderá vir a disponibilizar algumas ferramentas úteis no tratamento de dados, baseado em opções de frequência, que, neste momento, ainda não estão funcionais e, por isso, não são contempladas nas FAQ. Seria, por isso, um trabalho a desenvolver futuramente no sentido de melhorar as suas potencialidades.

2.2. APRESENTAÇÃO DAS FAQ

Apresentamos, seguidamente, as FAQ elaboradas para cada um dos *corpora* com a ressalva de que várias perguntas são comuns a ambos, ainda que com respostas diferentes em função das particularidades de cada um.

Quanto à ordem pela qual as questões são apresentadas e numeradas, esta obedece a um critério elementar: exemplificar um percurso de pesquisa do início ao fim.

Corpus de Português Língua Estrangeira	
Domínios de pesquisa	FAQ
I. Anotação do corpus	1. Quais as anotações disponíveis para cada produção escrita? 2. Como identificar os códigos dos estímulos a partir dos quais as produções escritas foram obtidas?
II. Construção de pesquisa	3. Como podem ser visualizados os resultados da pesquisa? 4. <i>Adicionar token</i> : como utilizar este critério de pesquisa? 5. Qual a funcionalidade de <i>Acrescentar ao lado</i> , no campo <i>construção de etiquetas</i> ? 6. Como pesquisar palavras contraídas? 7. Como pesquisar palavras hifenizadas? 8. Como pesquisar sufixos e prefixos? 9. Como pesquisar sequências de palavras de acordo com a ordem que ocupam na frase, combinando <i>Classe morfosintática / Lema / Forma do aluno</i> ? 10. Como devem ser pesquisados os sinais de pontuação? 11. Existe alguma forma de pesquisar palavras não coincidentes com a língua alvo no que se refere à ortografia ou à acentuação gráfica? 12. A pesquisa de dados de natureza pragmática é possível? 13. Como obter a frequência de determinadas ocorrências em função de diferentes critérios?
	14. Em que formatos podem ser descarregados os textos?

III. Armazenamento de dados	<p>15. Como aceder ao perfil dos informantes?</p> <p>16. Como descarregar a listagem de resultados?</p> <p>17. Como descarregar os dados de frequência de uma ocorrência?</p> <p>18. Como guardar e comparar expressões de pesquisa feitas anteriormente?</p>
------------------------------------	---

Tabela 4: Conjunto de 18 FAQ, distribuídas por domínios de pesquisa, referentes à pesquisa no PEAPL2_PLE.

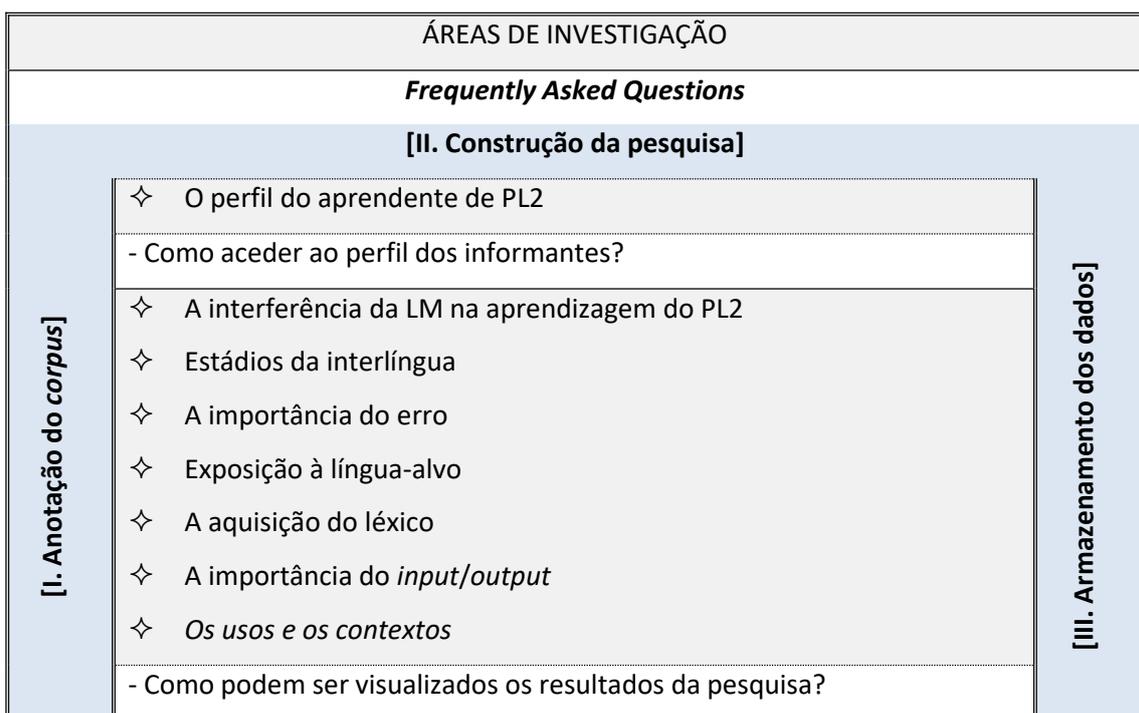
Corpus Oral de Português L2	
Domínios de pesquisa	FAQ
I. Anotação do corpus	<p>1. Quais as anotações disponíveis para cada produção escrita?</p> <p>2. Como identificar os códigos dos estímulos a partir dos quais as produções escritas foram obtidas?</p>
II. Construção de pesquisa	<p>3. Como podem ser visualizados os resultados da pesquisa?</p> <p>4. <i>Adicionar token</i>: como utilizar este critério de pesquisa?</p> <p>5. Qual a funcionalidade de <i>Acrescentar ao lado</i>, no campo <i>construção de etiquetas</i>?</p> <p>6. Como pesquisar palavras contraídas?</p> <p>7. Como pesquisar palavras hifenizadas?</p> <p>8. Como pesquisar sufixos e prefixos?</p> <p>9. Como pesquisar sequências de palavras de acordo com a ordem que ocupam na frase, combinando <i>Classe morfossintática / Lema / Forma do aluno</i>?</p> <p>10. Como pesquisar estrangeirismos?</p> <p>11. A pesquisa de dados de natureza pragmática é possível?</p> <p>12. Como obter a frequência de determinadas ocorrências em função de diferentes critérios?</p> <p>13. Quais os aspetos a considerar na pesquisa por tarefa?</p>

III. Armazenamento de dados	<p>14. Em que formatos podem ser descarregadas as produções?</p> <p>15. Como aceder ao perfil dos informantes?</p> <p>16. Como descarregar a listagem de resultados?</p> <p>17. Como descarregar os dados de frequência de uma ocorrência?</p> <p>18. Como guardar e comparar expressões de pesquisa feitas anteriormente?</p>
------------------------------------	--

Tabela 5: Conjunto de 18 FAQ, distribuídas por domínios de pesquisa, referentes à pesquisa no CORal-Co.

As FAQ que diretamente se relacionam com temas de investigação recolhidos na primeira fase do projeto são as do grupo II, referentes à construção de pesquisa. As restantes questões são transversais a todos os domínios de “construção de pesquisa”.

Embora o foco da investigação se centre num determinado tema, por vezes, as áreas de pesquisa não são estanques, na medida em que a descrição de uma estrutura da língua envolve outras áreas afins. Ainda assim, elaborámos um quadro-síntese com as FAQ e as áreas de investigação onde estas poderão ter aplicabilidade mais imediata.



	<ul style="list-style-type: none"> - Existe alguma forma de pesquisar palavras não coincidentes com a língua-alvo no que se refere à ortografia ou à acentuação gráfica? - Como obter a frequência de determinadas ocorrências em função de diferentes critérios? - A pesquisa de dados de natureza pragmática é possível? - Como pesquisar estrangeirismos? 	
	<ul style="list-style-type: none"> ✧ Áreas críticas de aprendizagem da língua-alvo nos diferentes domínios (morfologia, sintaxe, fonética, semântica, pragmática) ✧ Padrões de aquisição ✧ Estudos de análise contrastiva 	
	<ul style="list-style-type: none"> - <i>Adicionar token</i>: como utilizar este critério de pesquisa? - Qual a funcionalidade de <i>A acrescentar ao lado</i>, no campo <i>construção de etiquetas</i>? - Como pesquisar palavras contraídas? - Como pesquisar palavras hifenizadas? - Como pesquisar sufixos e prefixos? - Como pesquisar sequências de palavras de acordo com a ordem que ocupam na frase, combinando <i>Classe morfossintática / Lema / Forma do aluno</i>? - Como devem ser pesquisados os sinais de pontuação? 	
	<p><i>A expressão escrita na aula de PLE</i></p> <p><i>Múltiplos olhares sobre Portugal, o português e os portugueses</i></p> <p><i>A aquisição da leitura</i></p> <p><i>A Entrevista oral inicial como instrumento de trabalho em PLNM</i></p> <p><i>A competência oral: uma abordagem por tarefas</i></p>	
	<ul style="list-style-type: none"> - Como identificar os códigos dos estímulos a partir dos quais as produções escritas foram obtidas? - Quais os aspetos a considerar na pesquisa por tarefa? 	

Tabela 6: FAQ distribuídas por áreas de investigação e por domínio de pesquisa .

2.3. CONSTRUÇÃO DAS RESPOSTAS E DISPONIBILIZAÇÃO NA PLATAFORMA

As respostas às FAQ estão redigidas numa linguagem clara e objetiva, a pensar no utilizador, e com imagens ilustrativas do *passo a passo*. Inicialmente, cada uma das respostas foi gravada em ficheiros individuais, em *PDF* (ANEXOS II e III) de onde consta a FAQ e a resposta ilustrada com a orientação de pesquisa, como se pode observar, a partir do exemplo apresentado.

2. Como podem ser visualizados os resultados de pesquisa? Há opções.

Os resultados da pesquisa podem ser visualizados de duas formas diferentes.

- Preencher os campos da construção da pesquisa, e, de seguida, clicar em **opções**.
- i. Os campos das **Opções de busca** estão pré-preenchidos por defeito porque os resultados são, por definição, apresentados em linha de contexto (*Key Word in Context*), correspondendo à combinação mais curta de 5 palavras.
- Quando uma pesquisa é criada, os resultados são sempre apresentados desta forma, automaticamente, não sendo necessário preencher qualquer campo.

The screenshot displays the 'Pesquisa no corpus' interface. At the top, there is a search bar labeled 'OQP Query' with a 'Pesquisar' button and a link to 'construção de pesquisa | ver | opções'. Below this is a section titled 'Opções de busca' with the following settings: 'Tipo de representação visual: * KWIC | Context', 'Tamanho do contexto: 5 * palavras', 'Ordenar por: Relevância', and 'Estratégia de combinação: Continuação mais longa'. The main section is 'Construção de pesquisa', which is divided into two columns: 'Pesquisa de texto' and 'Pesquisa do documento'. The 'Pesquisa de texto' column includes fields for 'Forma de abstr.' (set to 'graf a'), 'Forma corrigida' (set to 'graf a'), 'Classe morfológica' (set to 'construção de palavras'), and 'Linha' (set to 'opção'). The 'Pesquisa do documento' column includes fields for 'Racionalidade' (set to 'intencional'), 'Ligação externa' (set to 'intencional'), 'Problema' (set to 'intencional'), 'Fase de recolha' (set to 'intencional'), and 'Estado' (set to 'intencional'). At the bottom of the 'Construção de pesquisa' section, there are buttons for 'Adicionar token', 'Clicar pesquisa', 'Cancelar', and 'ajuda'.

Figura 1: Exemplo de resposta elaborada para a FAQ2, em formato *PDF*, para orientação de pesquisa no PEAPL2_PLE.

Posteriormente, as respostas foram gravadas em suporte audiovisual, na forma de vídeos explicativos, em suporte *mp4*. A gravação do registo áudio e da imagem são realizadas em simultâneo, não havendo recurso a suporte de leitura; desse facto resultam, naturalmente, pausas, hesitações ou reformulações, que, a nosso ver, não afetam a inteligibilidade do conteúdo e lhe conferem autenticidade.

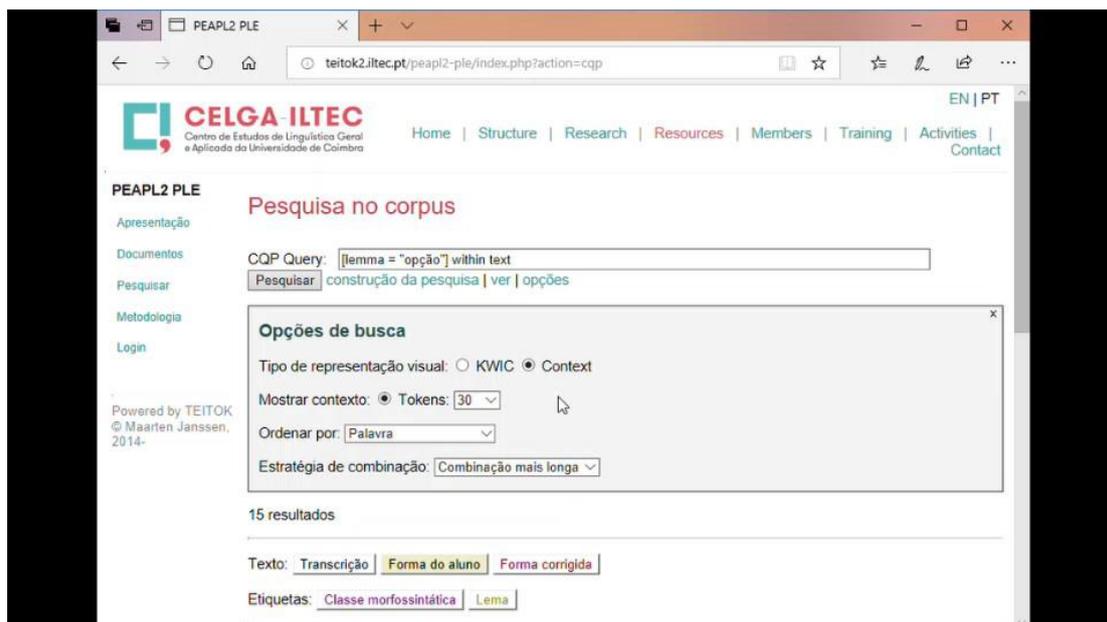


Figura 2: Fotograma de vídeo explicativo para a FAQ2, em formato *mp4*, para orientação de pesquisa no PEAPL2_PLE.

Numa fase prévia ao *upload* dos ficheiros na plataforma, houve necessidade de redigir um texto informativo introdutório à pesquisa e à consulta das FAQ com algumas orientações genéricas, mas necessárias para uma boa utilização das ferramentas de pesquisa, que a seguir se apresenta.

Pesquisa no corpus

CQP Query: [construção da pesquisa](#) | [ver](#) | [opções](#)

PEAPL2_PLE

[Ver todos os documentos](#)

- Os textos que constituem o Subcorpus Português Língua Estrangeira (PLE) estão armazenados num banco de dados em formato **TEI (Text Encoding Initiative)** para poderem ser pesquisáveis *online* através do sistema **CQL (Corpus WorkBench Query Language)**.
 - Para fazer uma pesquisa no *corpus* de PLE, preencha os campos disponíveis na **construção de pesquisa**.
- Antes de começar a pesquisa, saiba que:**
- A pesquisa simples, por palavra/expressão, sem preenchimento de campos na construção de pesquisa, devolve resultados relativos à **Forma do aluno**.
 - Sempre que construir uma pesquisa, deverá clicar primeiro em **Criar pesquisa**, para inserir os dados de pesquisa, e só depois em **Pesquisar**.
 - Se clicar em pesquisar, sem preencher nenhum campo, poderá visualizar uma listagem de **todos os documentos** disponíveis, identificados por nacionalidade, língua materna e nível de proficiência do aprendente, fase de recolha e estímulo.
 - Disponibilizamos um conjunto de **FAQ** cujas respostas orientam as pesquisas, não só no *corpus*, mas também na(s) página(s) do projeto.
- Nota:**
- Caso não esteja a conseguir obter resultados seguindo estas indicações, experimente mudar de **browser**.

Figura 3: Texto informativo introdutório à pesquisa e às FAQ, na página de pesquisa do PEAPL2_PLE.

Na última etapa, as FAQ foram disponibilizadas nas páginas de pesquisa dos respetivos *corpora*, na plataforma TEITOK, do seguinte modo: apresentação de cada uma das FAQ, inseridas nos domínios de pesquisa que definimos previamente, podendo a resposta ser consultada acedendo ao ficheiro *PDF* ou ao vídeo explicativo, como a seguir se ilustra.

Frequently Asked Questions (FAQ)

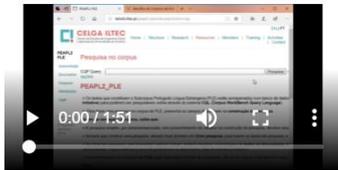
Para localizar mais rapidamente a informação que procura, as questões estão agrupadas de acordo com diferentes etapas de pesquisa.

I. ANOTAÇÃO DO CORPUS

II. CONSTRUÇÃO DA PESQUISA

III. ARMAZENAMENTO DE DADOS

FAQ14 Como aceder ao perfil dos informantes?



FAQ14_Informantes.pdf

Figura 4: Aspeto gráfico do *interface* de pesquisa após *upload* das FAQ na plataforma TEITOK.

Concluído o *upload* do texto informativo e das FAQ na plataforma TEITOk, nos dois suportes apresentados, o aspeto gráfico da página da pesquisa é o que a seguir se exhibe³³.

³³ A maioria dos exemplos apresentados relativamente à construção de resposta às FAQ e à sua disponibilização na plataforma TEITOK são referentes ao PEAPL2_PLE, uma vez que o COraL-Co ainda se encontra em desenvolvimento.

EN | PT

Home | Structure | Research | Resources | Members | Training | Activities | Contact

CELGA ILTEC
Centro de Estudos de Língua Geral
e Aplicados da Universidade de Coimbra

PEAPL2 PLE

Apresentação

Documentos

Pesquisar

Metodologia

Login

Pesquisa no corpus

CQP Query: [construção da pesquisa](#) | [ver](#) | [opções](#)

PEAPL2_PLE

[Ver todos os documentos](#)

Powered by TEITOK
© Maarten Janssen,
2014.

Os textos que constituem o Subcorpus Português Língua Estrangeira (PLE) estão armazenados num banco de dados em formato TEI (Text Encoding Initiative) para poderem ser pesquisáveis online através do sistema CQL (Corpus WorkBench Query Language).

Para fazer uma pesquisa no corpus de PLE, preencha os campos disponíveis na [construção de pesquisa](#).

Antes de começar a pesquisa, saiba que:

- A pesquisa simples, por palavra/expressão, sem preenchimento de campos na construção de pesquisa, devolve resultados relativos à *Forma do aluno*.
- Sempre que construir uma pesquisa, deverá clicar primeiro em **Crear pesquisa**, para inserir os dados de pesquisa, e só depois em **Pesquisar**.
- Se clicar em pesquisar, sem preencher nenhum campo, poderá visualizar uma listagem de todos os documentos disponíveis, identificados por nacionalidade, língua materna e nível de proficiência do aprendente, fase de recolha e estímulo.
- Disponibilizamos um conjunto de FAQ cujas respostas orientam as pesquisas, não só no corpus, mas também na(s) página(s) do projeto.

Nota:

- Caso não esteja a conseguir obter resultados seguindo estas indicações, experimente mudar de browser.

Frequently Asked Questions (FAQ)

Para localizar mais rapidamente a informação que procura, as questões estão agrupadas de acordo com diferentes etapas de pesquisa.

I. ANOTAÇÃO DO CORPUS

II. CONSTRUÇÃO DA PESQUISA

III. ARMAZENAMENTO DE DADOS

FAQ14 Como aceder ao perfil dos informantes?



[FAQ14_Informantes.pdf](#)

Figura 5: Aspeto gráfico da página de pesquisa do PEAPL2_PLE.

Estes materiais encontram-se disponíveis, sem restrições de acesso, na página de pesquisa de cada um dos *corpora*, designadamente:

<http://teitok2.iltec.pt/peapl2-ple/index.php?action=cqp> (PLE)

<http://teitok2.iltec.pt/coralco/index.php?action=cqp&act=advanced> (COral-Co)

Considerações finais sobre as FAQ na orientação de pesquisa de *corpora* de aprendentes

No início deste trabalho, elencámos algumas das vantagens dos *corpora* informatizados, entre elas, o facto de a pesquisa ser mais rápida e eficiente. Seria, por isso, contraproduativo se uma pesquisa, num qualquer *corpus*, se afigurasse morosa devido, por exemplo, a dúvidas de funcionamento da plataforma ou à complexidade da linguagem de pesquisa, e levasse o investigador a abandonar a sua pesquisa.

Uma vez que o conjunto de FAQ elaborado para o *corpus* PEAPL2_PLE e para o COraL-Co antecipa dificuldades que se possam colocar no momento da pesquisa, acreditamos que a sua disponibilização será uma mais-valia na realização de pesquisas, traduzindo-se em economia de tempo e no sucesso da pesquisa de dados. Nesta perspetiva, pensamos que as FAQ acrescentam valor à plataforma TEITOK e contribuem para a rentabilização dos *corpora* ao permitir aproximar o pesquisador do objeto de pesquisa, independentemente dos seus propósitos e interesses.

Acreditamos igualmente que este modelo de construção de FAQ, baseado nos objetivos inerentes à criação de cada *corpus*, poderá ser replicado por outros *corpora* informatizados em ambiente TEI. Se o caminho para chegar à informação não é suficientemente claro, é pertinente, a nosso ver, que o *corpus* se faça acompanhar de um conjunto de FAQ para orientar as pesquisas.

Responder às questões colocadas pelas FAQ significa não só ter testado as funcionalidades de pesquisa, mas também ter levado a cabo uma reflexão sobre a preparação do *corpus* e sobre as potencialidades do *software*. Ao longo deste processo, algumas questões que, à partida, poderiam parecer ser pertinentes para orientar a pesquisa deixam de o ser, precisamente porque é possível aperfeiçoar o sistema, anulando as dificuldades. Em última instância, este trabalho permite, então, melhorar o próprio sistema. A este propósito, aventura-se-nos a possibilidade de criar um sistema de FAQ, automático, com características de *Página de Ajuda*, comum a *corpora* de aprendentes desenvolvidos em ambiente TEI.

Ainda com vista à otimização da plataforma, gostaríamos de observar que, embora se tenha levado a cabo uma bateria de testes na plataforma, na ótica do utilizador, é sempre possível colocar-se uma questão, várias questões até, sobre casos de pesquisas que, até ao momento, não nos suscitaram qualquer dúvida ou dificuldade. A pensar nestes casos, uma das propostas a desenvolver futuramente recai sobre a criação de uma área reservada de utilizador que permita colocar questões relevantes, que integrem o conjunto de FAQ agora desenvolvido, no sentido de melhorar o funcionamento da área de pesquisa dos *corpora*. Inerente à criação de uma área reservada estaria

também o objetivo de criar, temporariamente, um histórico de pesquisas realizadas e guardar os dados obtidos.

BIBLIOGRAFIA/FONTES CONSULTADAS

- ANTHONY, L. (2013). *A critical look at software tools in corpus linguistics*. *Linguistic Research*, 30, 141-161.
- ANTUNES, S. *et al* (2016). *Apresentação do Corpus de Português Língua Estrangeira/Língua Segunda – COPLE2*. In *Revista Portuguesa de Linguística*, nº 1 – 10, 85-103.
- ARAÚJO, S. & TRABULO, P. (2014). *Da Linguística de corpus ao ensino/aprendizagem de Línguas: da teoria à prática*. In *Revista de Letras*, série III, nº 13, 7-21.
- DEL RIO, I., ANTUNES, S., MENDES, A. & JANSSEN, M. (2016). *Towards error annotation in a learner' corpus of portuguese*. In *Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition at SLTC, Umea, 16th November 2016*, 8–17. Linkoping University, Electronic Press.
- DEL RIO, I. & MENDES, A. (2018). *Error annotation in the COPLE2 corpus*. *Revista da Associação Portuguesa de Linguística*, nº 4 – 09/2018, 225-239.
- DÍAZ-NEGRILLO, A. & FERNÁNDEZ-DOMÍNGUEZ, J. (2006). *Error Tagging Systems for Learner Corpora*. *Revista española de lingüística aplicada*, Vol. 19, 83-102.
- FILLMORE, C. (1992). *"Corpus linguistics" or "Computer-aided armchair linguistics"*. In *Corpus Linguistics: Proceedings from a 1991 Nobel Symposium on Corpus Linguistics*, 35-66.
- FREITAS, C. (2015). *Corpus, Linguística Computacional e as Humanidades Digitais*. In LEITE, M. & GABRIEL, C. (orgs) *Corpus, Linguística Computacional e as Humanidades Digitais*. Rio de Janeiro, 23-56.
- GILQUIN, G. (2005). *From design to collection of learner corpora*. In Sylviane Granger, Gaëtanelle Gilquin & Fanny Meunier (eds). *The Cambridge Handbook of Learner Corpus Research*. Cambridge: Cambridge University Press, 9-34.
- GRANGER, S. (1998). *The Computer Learner Corpus: a Versatile New source of data for SLA research*. In Granger, S. (ed.) (1998). *Learner English on Computer*. Addison Wesley Longman: London & New York, 3-18.
- GRANGER, S. (2004). *Computer Learner Corpus Research: Current Status and Future Prospects*, 123-145.
- GRANGER, S., MEUNIER, F. & GILQUIN, G. (eds.) (2005), *The Cambridge Handbook of Learner Corpus Research*. Cambridge.

-
- GRANGER S. (2008). *Learner corpora*. In Lüdeling, A. & Kytö, M. (eds.) *Corpus Linguistics. An International Handbook*. Volume 1. Berlin & New York: Walter de Gruyter, 259-275.
- GRANGER, S. (2012). *How to use foreign and second language learner corpora*. In A. Mackey & S. Gass (eds.) *Research Methods in Second Language Acquisition: A Practical Guide*. Malden: Blackwell, 7-29.
- HARDIE, A. (2012). *CQPweb — combining power, flexibility and usability in a corpus analysis tool*. In *International Journal of Corpus Linguistics*, John Benjamins Publishing Company, 17:3, 380–409.
- HOFFMANN, S., & EVERT, S. (2006). *BNCweb (CQP edition) - the marriage of two corpus tools*. In S. Braun, K. Kohn, & J. Mukherjee (Eds.), *Corpus technology and language pedagogy?: new resources, new tools, new methods*. Frankfurt am Main: Peter Lang, 177-195.
- JANSSEN, M. (2016). *TEITOK: Text-Faithful Annotated Corpora*. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC), 4037-4043.
- JANSSEN, M. (2018). *TEITOK: TEI for corpus linguistics*.
- JANSSEN, M. (2019). *TEITOK as a Tool for Learner Corpora*. (Artigo não publicado)
- JOHNSON, J., NEWPORT, E. (1989). *Critical Period Effects in Second Language Learning: The Influence of Maturational State on the Acquisition of English as a Second Language*. In *COGNITIVE PSYCHOLOGY* 21, 60-99.
- MARTINS, C. (2008). *O papel diferenciado de subsistemas de memória de longo prazo nos processos de aquisição e de aprendizagem de uma L2*. A publicar in: Corrêa-Cardoso, João (Coord.). *A Linguagem na Pólis*. Coimbra (Centro de Estudos Clássicos e Humanísticos).
- McCARTHY, M. and O'KEEFFE, A. (2012). *Analysing Speech Corpora*. In T. Cobb (Ed.) *The Encyclopedia of Applied Linguistics*. New York: Wiley-Blackwell, 104-112.
- McENERY, T., XIAO, R. & TONO, Y. (2006). *Corpus-Based Language Studies An advanced resource book*. London & New York: Routledge.
- McENERY, T., & XIAO, R. (2010). *What corpora can offer in language teaching and learning*. In E. Hinkel (Ed.) *Handbook of Research in Second Language Teaching and Learning* (Vol. 2, 364-380). London & New York: Routledge.
- MENDES, Amália (2016). *The COPLE2 Corpus: a Learner Corpus for Portuguese*. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*. Portorož, Slovenia, European Language Resources Association (ELRA), 3207-32014.

-
- MENDES, Amália (2016). *Linguística de Corpus e outros usos dos corpora em linguística*. In Branco, António et al (2016) *A Língua Portuguesa na Era Digital / The Portuguese Language in the Digital Age*. White Paper Series. Berlin, Springer.
- MEUNIER, F. (2016). *Learner corpora and pedagogical applications*. In Murray, L. and Farr, F. (eds) 2016. *The Routledge Handbook of Language Learning and Technology*. Routledge.
- O'KEEFFE, A. & MCCARTHY, M. (2010). *Historical perspective: what are corpora and how have they evolved?* In Anne O'Keeffe & Michael McCarthy (eds.) *The Routledge Handbook of Corpus Linguistics*. London: Routledge, 3-13.
- NASCIMENTO, M. Fernanda Bacelar (2002). *O lugar do corpus na investigação linguística*. In A. Mendes & T. Freitas (orgs.) *Actas do XVIII Encontro da Associação Portuguesa de Linguística*. Porto: APL, 601-605.
- RAINERI, S. & DEBRAS, C. (2019). *Corpora and Representativeness Corpora and Representativeness: Where to go from now?*. *CogniTextes, Revue de l'Association française de linguistique cognitive*, vol. 19.
- SANTOS, I. (2016). *Corpus oral de PL2: um novo recurso para a investigação e ensino*. In *Revista da Associação Portuguesa de linguística*, nº 1 – 10, 745-760.
- SARMENTO, S. (2010). *Linguística de corpus: histórico, metodologia, campos de aplicação*. In *Revista Trama - Volume 6 - Número 12 - 2º Semestre*, 87 - 107.
- WINNE, M. (ed.) (2004). *Developing Linguistic Corpora: a Guide to Good Practice*. Oxford.
- SARDINHA, T. (2010). *Linguística de corpus: histórico e problemática*. *D.E.L.T.A.*, Vol. 16, Nº. 2, 323-367.
- SELINKER, L. (1972). *Interlanguage*. In *IRAL, International Review of Applied Linguistics in Language Teaching*, vol. 10:3, 209-231.
- SELINKER, L. (2014). *Interlanguage 40 years on: Three themes from here*. In HAN, Z. & TARONE, E. (eds) *Interlanguage. Forty years later*. *Language Learning & Language Teaching*, 39, John Benjamins, 221-246.

ANEXOS

ANEXO I - SISTEMA DE ETIQUETAS MORFOSSINTÁTICAS UTILIZADAS NO TEITOK

Sistema de etiquetas

pos = classe morfosintática

lemma = lema

form = forma do aluno

nform = forma corrigida (PEAPL2_PLE)

Word = forma corrigida (COraL-Co)

A	Adjetivo (Adj)	
1	Tipo	O ordinal Q qualificativo P possessivo
2	Grau	S superlativo V diminutivo
3	Género	F feminino M masculino C comum de dois géneros
4	Número	S singular P plural N invariável
5	Pessoa	1 1 ^a 2 2 ^a 3 3 ^a
6	Pessoa	S singular P plural N invariável

C	Conjunção													
1	Tipo	<table border="1"> <tr> <td>C</td> <td>coordenativa</td> </tr> <tr> <td>S</td> <td>subordinativa</td> </tr> </table>	C	coordenativa	S	subordinativa								
C	coordenativa													
S	subordinativa													
D	Determinante													
1	Tipo	<table border="1"> <tr> <td>A</td> <td>artigo</td> </tr> <tr> <td>D</td> <td>demonstrativo</td> </tr> <tr> <td>I</td> <td>indefinido</td> </tr> <tr> <td>T</td> <td>interrogativo</td> </tr> <tr> <td>N</td> <td>numeral</td> </tr> <tr> <td>P</td> <td>possessivo</td> </tr> </table>	A	artigo	D	demonstrativo	I	indefinido	T	interrogativo	N	numeral	P	possessivo
A	artigo													
D	demonstrativo													
I	indefinido													
T	interrogativo													
N	numeral													
P	possessivo													
2	Pessoa	<table border="1"> <tr> <td>1</td> <td>1ª</td> </tr> <tr> <td>2</td> <td>2ª</td> </tr> <tr> <td>3</td> <td>3ª</td> </tr> </table>	1	1ª	2	2ª	3	3ª						
1	1ª													
2	2ª													
3	3ª													
3	Género	<table border="1"> <tr> <td>F</td> <td>feminino</td> </tr> <tr> <td>M</td> <td>masculino</td> </tr> <tr> <td>C</td> <td>comum de dois géneros</td> </tr> </table>	F	feminino	M	masculino	C	comum de dois géneros						
F	feminino													
M	masculino													
C	comum de dois géneros													
4	Número	<table border="1"> <tr> <td>S</td> <td>singular</td> </tr> <tr> <td>P</td> <td>plural</td> </tr> <tr> <td>N</td> <td>invariável</td> </tr> </table>	S	singular	P	plural	N	invariável						
S	singular													
P	plural													
N	invariável													
5	Possessor	<table border="1"> <tr> <td>S</td> <td>singular</td> </tr> <tr> <td>P</td> <td>plural</td> </tr> </table>	S	singular	P	plural								
S	singular													
P	plural													
N	Nome													
1	Tipo	<table border="1"> <tr> <td>C</td> <td>comum</td> </tr> <tr> <td>P</td> <td>próprio</td> </tr> </table>	C	comum	P	próprio								
C	comum													
P	próprio													

2	Género	<table border="1"> <tr> <td>F</td> <td>feminino</td> </tr> <tr> <td>M</td> <td>masculino</td> </tr> <tr> <td>C</td> <td>comum de dois géneros</td> </tr> </table>	F	feminino	M	masculino	C	comum de dois géneros								
F	feminino															
M	masculino															
C	comum de dois géneros															
3	Número	<table border="1"> <tr> <td>S</td> <td>singular</td> </tr> <tr> <td>P</td> <td>plural</td> </tr> <tr> <td>N</td> <td>invariável</td> </tr> </table>	S	singular	P	plural	N	invariável								
S	singular															
P	plural															
N	invariável															
6	Grau	<table border="1"> <tr> <td>A</td> <td>aumentativo</td> </tr> <tr> <td>D</td> <td>diminutivo</td> </tr> </table>	A	aumentativo	D	diminutivo										
A	aumentativo															
D	diminutivo															
P	Pronome															
1	Tipo	<table border="1"> <tr> <td>D</td> <td>demonstrativo</td> </tr> <tr> <td>E</td> <td>exclamativo</td> </tr> <tr> <td>I</td> <td>indefinido</td> </tr> <tr> <td>T</td> <td>interrogativo</td> </tr> <tr> <td>N</td> <td>numeral</td> </tr> <tr> <td>P</td> <td>peçoal</td> </tr> <tr> <td>R</td> <td>relativo</td> </tr> </table>	D	demonstrativo	E	exclamativo	I	indefinido	T	interrogativo	N	numeral	P	peçoal	R	relativo
D	demonstrativo															
E	exclamativo															
I	indefinido															
T	interrogativo															
N	numeral															
P	peçoal															
R	relativo															
2	Pessoa	<table border="1"> <tr> <td>1</td> <td>1ª</td> </tr> <tr> <td>2</td> <td>2ª</td> </tr> <tr> <td>3</td> <td>3ª</td> </tr> </table>	1	1ª	2	2ª	3	3ª								
1	1ª															
2	2ª															
3	3ª															
3	Género	<table border="1"> <tr> <td>F</td> <td>feminino</td> </tr> <tr> <td>M</td> <td>masculino</td> </tr> <tr> <td>C</td> <td>invariável</td> </tr> </table>	F	feminino	M	masculino	C	invariável								
F	feminino															
M	masculino															
C	invariável															
4	Número	<table border="1"> <tr> <td>S</td> <td>singular</td> </tr> <tr> <td>P</td> <td>plural</td> </tr> </table>	S	singular	P	plural										
S	singular															
P	plural															

		N	invariável
5	caso	N	nominativo
		A	acusativo
		D	dativo
		O	oblíquo
R	Advérbio		
1	Tipo	N	negação
		G	outros
S	Adposição		
	{%}	SP	Preposição
1	Tipo	P	preposição
V	Verbo		
1	Tipo	M	principal
		A	auxiliar
		S	semiauxiliar
2	Modo	I	indicativo
		S	conjuntivo
		M	imperativo
		P	particípio passado
		G	gerúndio
		N	infinitivo
3	Tempo	P	presente
		I	pretérito imperfeito

		F	futuro
		S	pretérito perfeito
		C	condicional
		M	pretérito mais que perfeito
4	Pessoa	1	1ª
		2	2ª
		3	3ª
5	Número	S	singular
		P	plural
6	Género	F	feminino
		M	masculino
		C	comum
Z	Número		
1	Tipo	d	parte do todo
		m	moeda
		p	fração
		u	unidade
W	Data		
I	Interjeição		
F	Pontuação		
1	Tipo	a	ponto exclamação
		c	vírgula
		d	dois pontos
		e	aspas

2

	f	parêntesis
	g	hífen
	h	barra
	i	ponto interrogação
	l	chaveta
	p	ponto final
	q	parêntesis retos
	s	reticências
	t	percentagem
	x	ponto e vírgula
	z	outros
Posição	c	abrir
	t	fechar

**ANEXO II - FAQ E RESPOSTAS (PDF) DISPONÍVEIS NA PÁGINA
DE PESQUISA DO PEAPL2_PLE**

1. Quais as anotações disponíveis para cada produção escrita?

✍ A plataforma disponibiliza, para cada produção escrita, as seguintes anotações:

Opções de representação

Texto: Transcrição Forma do aluno Forma corrigida - Mostrar: Cores - Etiquetas: Classe morfosintática Lema

i. Transcrição

- ✍ A *Transcrição* corresponde ao texto original do aprendente, sem qualquer tipo de alteração ou correção por parte do professor, respeitando as suas hesitações, correções ou adições.
- ✍ Na fase inicial de transcrição dos textos, em formato *Word*, adotou-se um código de transcrição que sofreu pequenos reajustes aquando da transferência dos textos para o ambiente TEI.

Word	XML
< (...) > segmentos riscados ilegíveis	[...] segmentos riscados ilegíveis
< xxx > segmentos riscados	___ segmentos riscados
/ xxx / segmentos acrescentados	Segmentos acrescentados (cor)
xxxx nomes próprios que permitam identificar o informante.	Xxxx nomes próprios que permitam identificar o informante.
/* xxx / leituras conjeturadas	

(Em alternativa, consultar [aqui](#))

Em contraposição disso, eu estudo em Heidelberg [a](#) que é uma [...] cidade universitária, vital [e](#) ~~urbano~~ com um clima muito urbano. Enquanto nós ~~não~~ encontramos ~~na região da minha família~~ nenhuns monumentos importantes ~~na região da minha família~~ (excepções: um castello velho e uma cidade do século 17), é Heidelberg um ~~local~~ o lugar o qual os turistas conhecem bem.

ii. Forma do aluno

✍ É a versão final do texto escrito pelo aprendente.

Em contraposição disso, eu estudo em Heidelberg a que é uma [...] cidade universitária, vital e com um clima muito urbano. Enquanto nós não encontramos nenhuns monumentos importantes **na região da minha família** (excepções: um castello velho e uma cidade do século 17), é Heidelberg um o lugar o qual os turistas conhecem bem. Depois

iii. Forma corrigida

✍ Apenas se procedeu a correções de natureza ortográfica.

✍ Todos os desvios de natureza ortográfica que pudessem implicar desvios de natureza morfosintática ou semântica não foram tidos em conta para não interferir na leitura e interpretação dos dados.

✍ Os estrangeirismos bem como as leituras conjeturadas também não foram corrigidos sob pena de comportarem alterações de significado e/ou de intenção comunicativa por parte do aprendente difíceis de confirmar.

Opções de representação

Texto: Transcrição Forma do aluno Forma corrigida - Mostrar: Cores - Etiquetas: Classe morfosintática Lema

Eu gosto de fazer muito durante meus tempos livres, mas sobretudo, eu gosto de viajar. Durante, o **Verão** passado, fui para Índia e Tailândia com os meus amigos da minha universidade. Visitámos, a **Índia** durante onze dias, y visitámos **os cidades** de Mumbai, e New Delhi e uma aldeia que chama-se Puskhar. **Montámos** de camelos pelo deserto e **comemos** muito cariles e outra comida **exótica**. Visitámos muitos mercados e tudo fue muito muito barato! Comprei demais e o meu saco fue muito pesado.

iv. Cores

✍ Existem dois códigos de cores para o texto:

i. **Vermelho**: formas corrigidas;

ii. **Azul**: segmentos acrescentados.

Opções de representação

Texto: **Transcrição** **Forma do aluno** **Forma corrigida** - Mostrar: **Cores** - Etiquetas: **Classe morfosintática** **Lema**

Olá	XXXXX	!
olá	xxxxxx	!

Há	Quanto	tempo	não	te	vejo	!
háver	quanto	tempo	não	te	ver	!

Já	sei	,	falamos	não	há	muito	tempo	,	mas	estou	cheia	de	saudades	tuas	.	E	tenho	muitas	aventuras	novas	para	te	contar	.	A	primeira	e	mais
já	sei	,	falamos	não	há	muito	tempo	,	mas	estou	cheia	de	saudades	tuas	.	E	tenho	muitas	aventuras	novas	para	te	contar	.	A	primeira	e	mais

importante	é	:	já	sou	tia	!	!	A	minha	sobrinha	nasceu	ontem	.	o	parto	correu	muito	bem	.	e	mãe	e	filha	estão	perfeitas	de	saúde	.	Vou	visitá-las
importante	é	:	já	sou	tia	!	!	A	minha	sobrinha	nasceu	ontem	.	o	parto	correu	muito	bem	.	e	mãe	e	filha	estão	perfeitas	de	saúde	.	Vou	visitá-las

no	próximo	fim-de-semana	.	Estou	tão	contente	!	Irei	de	carro	,	comprá-lo	em	Maio	.	em	segunda	mão	e	até	agora	não	tem	dado	problemas	.
no	próximo	fim-de-semana	.	Estou	tão	contente	!	Irei	de	carro	,	comprá-lo	em	Maio	.	em	segunda	mão	e	até	agora	não	tem	dado	problemas	.

Espero	que	continue	assim	!
esperar	que	continuar	assim	!

2. Como identificar os códigos dos estímulos a partir dos quais as produções escritas foram obtidas?

Existem 9 estímulos cuja codificação se encontra descrita na *Metodologia* ([aqui](#)).

The screenshot shows a web interface titled "Construção de pesquisa" (Search Construction). It is divided into two main sections: "Pesquisa do texto" (Text Search) and "Pesquisa do documento" (Document Search).
Under "Pesquisa do texto", there are four rows of search criteria, each with a dropdown menu and a text input field:
- "Forma do aluno" (Student form) with a dropdown set to "igual a" (equal to) and an empty text box.
- "Forma corrigida" (Corrected form) with a dropdown set to "igual a" and an empty text box.
- "Classe morfosintática" (Morphosyntactic class) with a dropdown set to "construção de etiquetas" (tag construction) and an empty text box.
- "Lema" (Lemma) with a dropdown set to "igual a" and an empty text box.
Below these are buttons for "Acrescentar token" (Add token), "Create query", "cancelar" (cancel), and "AJUDA" (HELP).
Under "Pesquisa do documento", there are four rows of search criteria, each with a dropdown menu:
- "Nacionalidade" (Nationality) with a dropdown set to "[seleccionar]" (select).
- "Lingua materna" (Mother tongue) with a dropdown set to "[seleccionar]".
- "Proficiência" (Proficiency) with a dropdown set to "[seleccionar]".
- "Fase de recolha" (Collection phase) with a dropdown set to "[seleccionar]".
Below these is a section for "Estímulo" (Stimulus) with a dropdown menu that is open, showing a list of codes: 1.1A, 1.1A, 33.1J, 50.2L, 52.2L, 55.2M, 6.1B, 69.3Q, 75.3S, and 77.3T.
At the bottom of the interface, there is a text box for entering a search query in CQL format, with an example: "[lemma='casa'] [pos='A.*']".

3. Como podem ser visualizados os resultados de pesquisa? *Há opções.*

Os resultados da pesquisa podem ser visualizados de duas formas diferentes.

📌 Preencher os campos da construção da pesquisa, e, de seguida, clicar em **opções**.

i. Os campos das **Opções de busca** estão pré-preenchidos por defeito porque os resultados são, por definição, apresentados em linha de contexto (*Key Word in Context*), correspondendo à combinação mais curta de 5 palavras.

📌 Quando uma pesquisa é criada, os resultados são sempre apresentados desta forma, automaticamente, não sendo necessário preencher qualquer campo.

Pesquisa no corpus

CQP Query: [Pesquisar](#) [construção da pesquisa](#) | [ver](#) | [opções](#)

Opções de busca

Tipo de representação visual: KWIC Context

Tamanho do contexto: 5 palavras

Ordenar por: Palavra

Estratégia de combinação: Combinação mais longa

Construção de pesquisa

Pesquisa do texto	Pesquisa do documento
Forma do aluno: igual a	Nacionalidade: [selecionar]
Forma corrigida: igual a	Língua materna: [selecionar]
Classe morfosintática: construção de etiquetas	Proficiência: [selecionar]
Lema: igual a opção	Fase de recolha: [selecionar]
	Estímulo: [selecionar]

[Adicionar token](#)

[Criar pesquisa](#) [cancelar](#) | [ajuda](#)

Pesquisa no corpus

CQP Query: [lemma = "opção"] within text Pesquisar construção da pesquisa | ver | opções

15 resultados

Texto: Transcrição Forma do aluno Forma corrigida

Etiquetas: Classe morfosintática Lema

contexto	campo, embora seja uma	opção	interessante e paussível por
contexto	consegue aportar. Se não tivermos	opção	desta fugida nalguns
contexto	vai que a melhor	opção	para viver é uma
contexto	escolha e mais possibilidades da	opção	nos serviços, no
contexto	minha país a bicicleta e uma	opção	muito útil, porque
contexto	meramente que há nenhuma outra	opção	, seja por causa
contexto	perto da cidade é uma	opção	desejável
contexto	as zonas rurais são uma boa	opção	. Aí a pessoa
contexto	cinema também era uma	opção	muito boa para mim
contexto	que eu gosto das duas	opções	. É certeza que
contexto	porque há muitas possibilidades e	opções	entre as quais escolher
contexto	campo onde não há muitas outras	opções	. Todos os dias
contexto	. Em resumo, as duas	opções	oferecem modos de vida
contexto	, já tenho experimentado as várias	opções	e bem sei que
contexto	mais restaurantes porque não há muitas	opções	e se quiser jantar

ii. Opcionalmente, os resultados podem ser apresentados por contexto, numa combinação mais longa de (e até 100) *tokens*.

✎ Para isso, selecionar os campos **Context** e **Mostrar contexto** e escolher o número de *tokens* que pretende.

✎ Por fim, clicar em **Criar pesquisa** e **Pesquisar**.

Pesquisa no corpus

CQP Query: Pesquisar construção da pesquisa | ver | opções

Opções de busca

Tipo de representação visual: KWIC Context

Mostrar contexto: Tokens:

Ordenar por:

Estratégia de combinação:

Construção de pesquisa

Pesquisa do texto	Pesquisa do documento
Forma do aluno: <input type="text" value="igual a"/>	Nacionalidade: <input type="text" value="[selecionar]"/>
Forma corrigida: <input type="text" value="igual a"/>	Língua materna: <input type="text" value="[selecionar]"/>
Classe morfosintática: <input type="text" value="construção de etiquetas"/>	Proficiência: <input type="text" value="[selecionar]"/>
Lema: <input type="text" value="igual a"/> <input type="text" value="opção"/>	Fase de recolha: <input type="text" value="[selecionar]"/>
	Estímulo: <input type="text" value="[selecionar]"/>

Pesquisa no corpus

CQP Query: [lemma = "opção"] within text construção da pesquisa | ver | opções

15 resultados

Texto:

Etiquetas:

lúdica. Mas a questão é cidade a campo? Eu, e digo com certeza absoluta, não conseguiria viver todo o tempo no campo, embora seja uma opção interessante e pausável por um tempo limitado [...] como é no verão, ou algumas semanas de férias. As principais vantagens que eu vejo na vida no campo é a possibilidade de ter um espaço próprio, ou precisamos, ou precisamos, numa actividade contínua sendo as vezes desejados esses momentos de paz que uma vida no campo consegue aportar. Se não tivermos opção desta fugida nalguns momentos, e a escolha continua cidade-campo, eu sem dúvida alguma moraria numa cidade. O que tipo de cidade? A cidade enorme. O terrorismo causa-me muito medo, a poluição sonora provoca dores de cabeça, o "nevoeiro" cidade. Daí vai que a melhor opção para viver é uma cidade pequena que reúne as vantagens da cidade maior e do campo. Porém, as pessoas de uma metrópole têm oportunidade de ir para a cidade, estamos perto de tudo: escolas, centros de saúde, lojas e cinemas. Em consequência, temos mais escolha e mais possibilidades da opção nos serviços, no emprego e no lazer. Para além disso, vivendo num cidade, acabamos por ter mais mobilidade, uma rapidez e mais económico. O conforto é menos importante e depende também a distância, mas, mais importante a duração. Na minha país a bicicleta e uma opção muito útil, porque o terreno é muito plano. Aqui em Portugal, com muitas montanhas a bicicleta é menos habitual. Apesar a bicicleta e muito pratico comer e beber bem e adoram de contar contos e piadas simultaneamente. Pessoas das ambas culturas condusem os carros rapidamente e reduzem a velocidade [...] meramente que há nenhuma outra opção, seja por causa de uma placa (às vezes) ou seja por causa dum acidente. Uma última curiosidade. Em ambas culturas há pessoas mais, o campo e um cenário mais adequado para criar animais. É por tudo isto que ter uma casa no campo mas perto da cidade é uma opção desejável tem as medidas. Especialmente para pessoas que não aguentam bem o stress e as condições ou que simplesmente têm filhos pequenos e medo, as zonas rurais são uma boa opção. A pessoa é capaz de relaxar melhor, manter um estilo de vida mais tranquilo e encontrar a felicidade em si. Normalmente a natureza faz bem às pessoas e bebendo um café ou um chá ou jogar junto com a família ou com os meus amigos. Ver um filme em casa ou no cinema também era uma opção muito boa para mim e gosto muito de fazer isso nos dias frios ou com tempo desagradável. A noite, na hora do jantar, mas quando era pequena costumava passar as férias de verão na casa do meu avô no campo. É por isso que eu gosto das duas opções. É certeza que temos vantagens e desvantagens nos dois casos, por exemplo viver na cidade é bom porque sempre temos tudo à disposição, e campo, são bons para viver, mas em diferentes momentos da vida. Eu sempre morei na cidade. Gosto dela porque há muitas possibilidades e opções entre as quais escolher. Eu adoro a gente. Adoro socializar com as pessoas e mais do que só gostar, preciso mesmo acordar de manhã e ir trabalhar, prefiro viajar de carro. Na minha situação o carro é o meio de transporte mais fácil porque vivo no campo onde não há muitas outras opções. Todos os dias viajo de casa para Coimbra mas, porque é difícil encontrar lugares de estacionamento, estaciono o carro no parque Ecovia Casa do Sal.

- Após a exibição dos resultados, passar com o curso por cima das palavras para uma rápida observação da etiquetagem das palavras.
- Para consultar a anotação linguística detalhada, aceder à produção do aprendente na íntegra, clicando em **contexto**, no início de cada linha de ocorrência.

Pesquisa no corpus

CQP Query: [lemma = "opção"] within text construção da pesquisa | ver | opções

15 resultados

Texto:

Etiquetas:

lúdica. Mas a questão é cidade a campo? Eu, e digo com certeza absoluta, não conseguiria viver todo o tempo no campo, embora seja uma opção interessante e pausável por um tempo limitado [...] como é no verão, ou algumas semanas de férias. As principais vantagens que eu vejo na vida no campo é a possibilidade de ter um espaço próprio, ou precisamos, ou precisamos, numa actividade contínua sendo as vezes desejados esses momentos de paz que uma vida no campo consegue aportar. Se não tivermos opção desta fugida nalguns momentos, e a escolha continua cidade-campo, eu sem dúvida alguma moraria numa cidade. O que tipo de cidade? A cidade enorme. O terrorismo causa-me muito medo, a poluição sonora provoca dores de cabeça, o "nevoeiro" cidade. Daí vai que a melhor opção para viver é uma cidade pequena que reúne as vantagens da cidade maior e do campo. Porém, as pessoas de uma metrópole têm oportunidade de ir para a cidade, estamos perto de tudo: escolas, centros de saúde, lojas e cinemas. Em consequência, temos mais escolha e mais possibilidades da opção nos serviços, no emprego e no lazer. Para além disso, vivendo num cidade, acabamos por ter mais mobilidade, uma rapidez e mais económico. O conforto é menos importante e depende também a distância, mas, mais importante a duração. Na minha país a bicicleta e uma opção muito útil, porque o terreno é muito plano. Aqui em Portugal, com muitas montanhas a bicicleta é menos habitual. Apesar a bicicleta e muito pratico comer e beber bem e adoram de contar contos e piadas simultaneamente. Pessoas das ambas culturas condusem os carros rapidamente e reduzem a velocidade [...] meramente que há nenhuma outra opção, seja por causa de uma placa (às vezes) ou seja por causa dum acidente. Uma última curiosidade. Em ambas culturas há pessoas

férias	Nome (NCFP000)	comum, feminino, plural
férias	Classe morfosintática	o
férias	Lema	férias

4. Adicionar *token*: como utilizar este critério de pesquisa? *É muito fácil!*

- ✍ A cada *token* corresponde um lema.
- ✍ Para pesquisar uma determinada expressão, constituída por mais do que uma palavra, escrever no campo **Lema** a primeira palavra da sequência e clicar em **Adicionar token**.
- ✍ Repetir o processo adicionando outros lemas/*tokens*, sempre por ordem, até obter a expressão pretendida.
- ✍ Sempre que adiciona um *token*, pode observar-se o conjunto de lemas adicionados pela ordem em que serão pesquisados.

Pesquisa no corpus

CQP Query: Pesquisar construção da pesquisa | ver | opções

Construção de pesquisa

1	Lema = ser	2	Lema = muito	3	Lema = fácil
---	------------	---	--------------	---	--------------

Pesquisa do texto

Forma do aluno	igual a	<input type="text"/>
Forma corrigida	igual a	<input type="text"/>
Classe morfosintática	construção de etiquetas	<input type="text"/>
Lema	igual a	<input type="text"/>

Adicionar token

Criar pesquisa cancelar | ajuda

Pesquisa do documento

Nacionalidade	[selecionar]
Língua materna	[selecionar]
Proficiência	[selecionar]
Fase de recolha	[selecionar]
Estímulo	[selecionar]

Pesquisa no corpus

CQP Query: [lemma = "ser"] [lemma = "muito"] [lemma = "fácil"] within text Pesquisar construção da pesquisa | ver | opções

4 resultados

Texto: Transcrição Forma do aluno Forma corrigida

Etiquetas: Classe morfosintática Lema

contexto	distante du minha casa. Não	e muito facil	de trabalhar efectivo por
contexto	aulas vão bastante boas – não	são muito facéis	compreender, mas acho
contexto	quero e tem aeroportos por isso	é muito facil	e conveniente para viajar
contexto	tudo o ano. A razão	é muito fácil	; morar no

- ✍ É possível utilizar a opção **Adicionar token** sempre que são preenchidos os campos **Lema**, **Forma corrigida** ou **Forma do aluno**, mas não é possível fazê-lo com o campo **Classe morfosintática**.

- 📌 Apenas se pode combinar a opção **Adicionar token** com o campo **Classe morfossintática** no final da expressão pretendida, isto é, primeiro adicionam-se todos os *tokens* pretendidos e, por último, pode adicionar-se a classe morfossintática à qual pertence a palavra que surgirá a seguir à sequência de *tokens*.

Pesquisa no corpus

CQP Query: construção da pesquisa | ver | opções

Construção de pesquisa

1	Lema = fazer	2	Lema = se
---	--------------	---	-----------

Pesquisa do texto

Forma do aluno	igual a	<input type="text"/>
Forma corrigida	igual a	<input type="text"/>
Classe morfossintática	construção de etiquetas	N.*
Lema	igual a	<input type="text"/>

Pesquisa do documento

Nacionalidade	[selecionar]
Língua materna	[selecionar]
Proficiência	[selecionar]
Fase de recolha	[selecionar]
Estímulo	[selecionar]

Pesquisa no corpus

CQP Query: construção da pesquisa | ver | opções

1 resultados

Texto:

Etiquetas:

contexto centro proprio do bairro onde **fazem-se festas** e juntam-se para

[Descarregar resultados](#) - [Memorizar expressão de busca](#)

- 📌 **Nota:** Quando combinamos o preenchimento do campo **Lema** com o campo **Classe morfossintática**, estamos a pesquisar uma palavra específica que pertence a uma determinada classe.

Pesquisa no corpus

CQP Query: [construção da pesquisa](#) | [ver](#) | [opções](#)

Construção de etiquetas: Classe morfossintática

POS principal: Adverbo
[selecionar]
 Adjetivo
Adverbo

Construção

Pesquisa de

Forma Pontuação a

Forma Preposição a

Forma Pronome

Classe morfo Verbo

Lema Igual a muito

Pesquisa do documento

Nacionalidade [selecionar]

Lingua materna [selecionar]

Proficiência [selecionar]

Fase de recolha [selecionar]

Estímulo [selecionar]

Pesquisa no corpus

CQP Query: [pos = "R." & lemma = "muito"] within text [construção da pesquisa](#) | [ver](#) | [opções](#)

1857 resultados • A mostrar 0 - 100 ([seguintes](#))

Texto:

Etiquetas:

contexto	DE JULHO SOU UMA PESSOA	MUITO	AMBICIOSA, SIMPÁTICA E SOCIÁVEL
contexto	VERDE. TENHO UMA FAMILIA	MUITO	ESPECIAL. OS MEUS PAIS
contexto	OS MEUS PAIS SÃO PESSOAS	MUITO	BOAS, TRABALHADORES E HONESTAS
contexto	TEM 18 ANOS. GOSTEI	MUITO	DESTA EXPERIÊNCIA EM COIMBRA
contexto	A SAUDE DAS PESSOAS	MUITO	IMPORTANTES PARA MIM, E
contexto	DE PODER ACHAR UM TRABALHO	MUITO	BOM PARA O MEU FUTURO
contexto	COMO A MINHA VIDA É	MUITO	DIFERENTE DO QUE ERA
contexto	SEJA QUE AS ESCOLAS FICAM	MUITO	LONGE DE AQUI E SEM
contexto	CIUDADE SOPPORTÁVEL QUE NÃO TRAZ	MUITO	STRESS [→] ÀS PESSOAS QUE
contexto	QUAL TINHA PASADO	MUITO	TEMPO PARA AJUDAR A CORTAR
contexto	MINHA CIUDADE (NÃO	MUITO	GRANDE E CAOTICA) E
contexto	MENO 10 ANOS [→] NÃO [→] GOSTAVA	MUITO	DA VIDA NÓ
contexto	DE VISTA HISTÓRICCO, E	MUITO	AGRADÁVEL É NÃO E QUASI
contexto	VERDADE É QUE ISSO DEPENDE	MUITO	RESPEITO A DIMENSÃO DA
contexto	. OU SEJA, É	MUITO	PROVÁVEL QUE QUALQUER NÃO GOSTERIA
contexto	TODAS AS VEZES EM NEGOCIOS	MUITO	LONGE DAS SUAS CASA E
contexto	ENQUENTADOR ELECTRICOS, QUE SÃO	MUITO	[→] CUSTOSOS. ALGUMAS CASAS SÃO

- 📖 Já quando queremos pesquisar uma palavra seguida de outra que pertence a determinada classe morfossintática, então, é necessário, depois de preencher o campo **Lema**, **Adicionar token** e só depois escolher a **Classe morfossintática**.

Pesquisa no corpus

CQP Query: [Pesquisar](#) [construção da pesquisa](#) | [ver](#) | [opções](#)

Construção de etiquetas: Classe morfosintática

POS principal: [selecionar]

Tipo: [qualquer]

Inserir Acres

Construção de etiquetas

1 Lema = *multo*

Pesquisa do texto

Forma do aluno:

Forma corrigida:

Classe morfosintática:

Lema:

Adicionar token

Críar pesquisa cancelar ajuda

Pesquisa do documento

Nacionalidade:

Língua materna:

Proficiência:

Fase de recolha:

Estímulo:

Pesquisa no corpus

CQP Query: [Pesquisar](#) [construção da pesquisa](#) | [ver](#) | [opções](#)

191 resultados • A mostrar 0 - 100 (seguintes)

Texto:

Etiquetas:

contexto	SEJA QUE AS ESCOLAS FIGAM	MUITO LONGE	DE AQUI E SEM [] UM
contexto	TODAS AS VEZES EM NEGOCIOS	MUITO LONGE	DAS SUAS CASA E COM
contexto	discótecas [] ali e não há	multas aqui	em Coimbra. Eu gosto
contexto	cidade porque acho que tem	multas mais	vantagens que desvantagens. Numa
contexto	casa. Ela acolheu-me	multo bem	diz-me de fazer
contexto	: "Kota, jogas	multo bem!!"	Fiquei tanto
contexto	isto não gosto de ler	multo agora	. Em vez de ler
contexto	falaram e tive que praticar	multo antes	de ser capado de ler
contexto	muito bom. Eu gosto	multo aqui	. Um beijinho XXXXX.
contexto	os meus colegas ajudaram-me	multo até	consegui escolher o tema do
contexto	dia maravilhoso. Eu fico	multo bem	. Sou estudante de Língua
contexto	. Na república estou	multo bem	e gosto das mininas
contexto	ontem, o parto correu	multo bem	, e mãe e filha
contexto	. Eu acho que te	multo bem	. Mas a minha prima
contexto	família? Por aqui estou	multo bem	Coimbra é fabulosamente,
contexto	. Mas agora eu estôu	multo bem	, e esta cidade gosto
contexto	de reconhecer que não toco	multo bem	. disfruto do ambiente

5. Qual a funcionalidade de *A acrescentar ao lado*, na *Construção de etiquetas*? *Tenho aulas e tenho aprendido muito.*

A opção **A acrescentar ao lado** permite fazer uma pesquisa inserindo uma classe morfosintática em alternativa à que foi inserida previamente, ou seja, permite pesquisar a ocorrência de diferentes classes em determinado contexto.

- ✎ Consideremos o exemplo do verbo *ter*: poder ser um verbo auxiliar ou um verbo pleno dependendo do contexto frásico em que ocorre.
- ✎ Esta opção vai permitir pesquisar, simultaneamente, dois dos contextos em que o verbo pode ocorrer, por exemplo, *ter+verbo // ter+nome*.
- ✎ Primeiro, preencha o campo **Lema** com a forma do infinitivo do verbo *ter* e clique em **Adicionar token**.

Pesquisa no corpus

CQP Query: construção da pesquisa | ver | opções

Construção de pesquisa

Lema = *ter*

Pesquisa do texto

Forma do aluno	igual a	<input type="text"/>
Forma corrigida	igual a	<input type="text"/>
Classe morfosintática	construção de etiquetas	<input type="text"/>
Lema	igual a	<input type="text"/>

cancelar | ajuda

Pesquisa do documento

Nacionalidade	[selecionar]
Lingua materna	[selecionar]
Proficiência	[selecionar]
Fase de recolha	[selecionar]
Estimulo	[selecionar]

- ✎ De seguida, clique em **construção de etiquetas > POS principal**, seleccione a categoria **Nome** e clique em **Inserir**.

Construção de etiquetas: Classe morfosintática

POS principal: Nome

Tipo [qualquer]

Gênero [qualquer]

Número [qualquer]

Grau [qualquer]

Inserir Acrescentar ao lado

Construção de pesquisa

Lema = ter

Pesquisa do texto

Forma do aluno igual a

Forma corrigida igual a

Classe morfosintática construção de etiquetas

Lema igual a

Adicionar token

Criar pesquisa cancelar ajuda

Pesquisa do documento

Nacionalidade [selecionar]

Língua materna [selecionar]

Proficiência [selecionar]

Fase de recolha [selecionar]

Estímulo [selecionar]

Pesquisa no corpus

CQP Query: construção da pesquisa | ver | opções

Construção de pesquisa

Lema = ter

Pesquisa do texto

Forma do aluno igual a

Forma corrigida igual a

Classe morfosintática construção de etiquetas N.*

Lema igual a

Adicionar token

Criar pesquisa cancelar ajuda

Pesquisa do documento

Nacionalidade [selecionar]

Língua materna [selecionar]

Proficiência [selecionar]

Fase de recolha [selecionar]

Estímulo [selecionar]

- 📄 Volte a **construção de etiquetas** > **POS principal**, selecione a categoria **Verbo** e clique em **Acrescentar ao lado**. A pesquisa está construída.
- 📄 Finalize clicando em **Criar pesquisa** e **Pesquisar** para obter os resultados.

Pesquisa no corpus

CQP Query: construção da pesquisa | ver | opções

Construção de pesquisa

Lema = ter

Pesquisa do texto

Forma do aluno: igual a

Forma corrigida: igual a

Classe morfosintática: construção de etiquetas N.*[V.*]

Lema: igual a

Pesquisa do documento

Nacionalidade:

Língua materna:

Proficiência:

Fase de recolha:

Estímulo:

As estruturas *ter+nome* /*ter+verbo* surgem destacadas no contexto em que ocorrem.

Pesquisa no corpus

CQP Query: construção da pesquisa | ver | opções

585 resultados • A mostrar 0 - 100 (seguintes)

Texto:

Etiquetas:

contexto	MESMO, POR ISSO O BAIRRO	TEM PREDIOS	E RUAS NA MAIOR
contexto	EU SOU ALTA,	TENHO CABELO	CASTANHO, ONDULADO, OLHOS
contexto	CONSEGUI VOAR [...] PORQUE OU NÃO	TENHO TEMPO	, OU CHOVE? QUE PENA
contexto	PEQUEN HORTO NO QUAL	TINHA PASADO	MUITO TEMPO PARA AJUDAR A
contexto	farto desta situação	Tem coisas	más. No meu
contexto	altíssimas e também planície.	Tem fronteiras	com tres países e cada
contexto	que faz o vizinho.	Tem lados	bons e más, porque
contexto	vida é bastante conveniente.	Temos Internet	, bar, restaurante e
contexto	dividido em 5 regiões.	Temos West-Vlaanderen	("O Flandres Occidental")
contexto	aqui estou todos bem.	Temos escrever	os testes e houve de
contexto	um quarto pra mim.	Tendo feito	isso fui varias vezes mas
contexto		Tendo morado	várias vezes no estrangeiro
contexto	altura que nós desejamos.	Tenho andado	muito ocupada com os exames
contexto	personas de culturas diferentes.	Tenho aprendido	, que a mentalidade e
contexto	— o Rio Mondego.	Tenho aulas	todos os dias desde as
contexto	embora não falem inglês!	Tenho aulas	todos os dias, e
contexto	dia 5 de Setembro.	Tenho aulas	todos os dias [...] e tenho

Nota: Este tipo de pesquisa também pode ser útil no estudo de casos relacionados com (in)transitividade dos verbos, regência preposicional, colocações na frase.

6. Como pesquisar palavras contraídas? *Deste e daquele.*

As palavras contraídas correspondem a dois lemas pertencentes a classes morfossintáticas distintas.

Existem duas formas de pesquisar palavras contraídas:

i. Por **Lema**, obtendo todas as palavras que resultam da contração com o lema pesquisado;

ii. Por combinação de dois lemas, no caso de pretender pesquisar uma preposição específica contraída com uma palavra pertencente a uma determinada classe /subclasse.

Pesquisa no corpus

CQP Query: [construção da pesquisa](#) | [ver](#) | [opções](#)

Construção de pesquisa

1 Lema = de 2 Lema = este

Pesquisa do texto

Forma do aluno igual a

Forma corrigida igual a

Classe morfossintática construção de etiquetas

Lema igual a

[ajuda](#)

Pesquisa do documento

Nacionalidade [selecionar]

Língua materna [selecionar]

Proficiência [selecionar]

Fase de recolha [selecionar]

Estímulo [selecionar]

Pesquisa no corpus

CQP Query: [construção da pesquisa](#) | [ver](#) | [opções](#)

119 resultados • a mostrar 0 - 100 (seguintes)

Texto:

Etiquetas:

contexto	Mas a maneira de viajar	destes	meios era para mim sempre
contexto	nossa capital em comboio.	Desta	altura o comboio é synónimo
contexto	Portugal. Eu gosto muito	deste	país e sobretudo da
contexto	pontos mais altos	desta	serra tão bonita. Isso
contexto	mais gosto é poder disfrutar	desta	esperiencia com os meus amigos
contexto	para que? Porque gosto	destas	coisas? Só há uma
contexto	minhas amigas de Coimbra.	Desta	forma pode aprender distintas formas
contexto	. Vou ter muitos saudades	deste	país, mas estou feliz
contexto	são a tranquilidade e serenidade	desta	. Pelo contrário numa
contexto	. Se não tivermos opção	desta	fugida alguns momentos,
contexto	Vou escrever de cada um	destes	no ordem descente dos
contexto	18 ANOS. GOSTEI MUITO	DESTA	EXPERIÊNCIA EM COIMBRA, É
contexto	razão, eu gosto muito	desta	vida sem stress e sem
contexto	Coimbra e curiosamente gosto bastante	desta	cidade pequena. Embora tenha

7. Como pesquisar palavras hifenizadas? *As pessoas juntam-se ao fim de semana.*

Como o hífen não foi lematizado, são duas as opções de pesquisa possíveis para os dois tipos de palavras hifenizadas:

i. palavras compostas: são consideradas como apenas um lema e, por isso, devem ser pesquisadas por **Lema**.

Pesquisa no corpus

CQP Query: [lemma = "fim-de-semana"] within text [construção da pesquisa](#) | [ver](#) | [opções](#)

58 resultados

Texto:

Etiquetas:

contexto JARDINES E PRAÇAS ONDE PASSEAR OS **FINS DE SEMANAS** E OS DIAS QUENTES DO
contexto tenho férias o no fim **de semana** . Quando vou ter
contexto mais lento. | Eu vou cada **fim de semana** parà Palermo para o
contexto Quando eu era pequena íamos para **fim de semana**
contexto calme de noite e no **fim de semana**
contexto são muito simpáticos. No **fim de semana**
contexto anos de adolescência, buscando cada **fim de semana**
contexto campo. Muitos dos **fim de semana**
contexto e para o mar durante o **fim de semana**
contexto viajar para O Porto no **fim de semana**
contexto até meia noite, e o **fim de semana** de três da
contexto fazemos outras coisas. o **fim de semana** fomos ao Serra
contexto e depois para Lisboa no **fim de semana** . Durante a semana
contexto). | Geralmente, ao **fim de semana** , não tenho outras
contexto "Technologia". Ao **fim de semana** , normalmente, jogo
contexto amigos, mas na ultima **fim de semana** nos fomos para a
contexto o meu português. Nos **fim de semana** muitas vezes não como
contexto , acostumo a ir no **fim de semana** ou no verão
contexto muitos trabalhos práticos. No **fim de semana** gosto muito sair com

Forma corrigida	fim-de-semana
Classe morfosintática	Nome (NCMS000) comum; masculino; singular
Lema	fim-de-semana

ii. conjugação pronominal: deve considerar-se dois lemas (o verbo e o clítico).

Pesquisa no corpus

CQP Query: [lemma = "juntar"] [lemma = "se"] within text [construção da pesquisa](#) | [ver](#) | [opções](#)

3 resultados

Etiquetas:

contexto bairro onde fazem-se festas e **juntam-se** para actividades culturais.
contexto meia noite. As famílias **juntam-se** com amigos, fazem
contexto estão a dançar, o povo **junta-se** para lanchar e ouvir

Nota: A pesquisa por **Pontuação > hífen** devolve, por defeito, ocorrências com travessão e não com hífen.

8. Como pesquisar sufixos e prefixos? *É possível!*

- ✎ Existe uma forma de pesquisar sufixos e prefixos que, não sendo exclusiva para este fim, é bastante produtiva.
- ✎ Ao preencher o campo **Lema**, selecionar primeiro **começa por / termina com** e depois escrever o prefixo / sufixo pretendidos.
- ✎ Da mesma forma, pode pesquisar um radical, selecionando a opção **contém** no **Lema**.
- ✎ O preenchimento deste campo pode ser combinado com o campo **Classe morfosintática** para refinar a pesquisa.

Pesquisa no corpus

CQP Query: [construção da pesquisa](#) | [ver](#) | [opções](#)

Construção de pesquisa

Pesquisa do texto		Pesquisa do documento	
Forma do aluno	<input type="text" value="igual a"/>	Nacionalidade	<input type="text" value="[selecionar]"/>
Forma corrigida	<input type="text" value="igual a"/>	Língua materna	<input type="text" value="[selecionar]"/>
Classe morfosintática	<input type="text" value="construção de etiquetas"/>	Proficiência	<input type="text" value="[selecionar]"/>
Lema	<input type="text" value="termina em"/>	Fase de recolha	<input type="text" value="[selecionar]"/>
	<input type="text" value="AQ.*"/>	Estímulo	<input type="text" value="[selecionar]"/>
	<input type="text" value="vel"/>		

[cancelar](#) | [ajuda](#)

Pesquisa no corpus

CQP Query: [construção da pesquisa](#) | [ver](#) | [opções](#)

325 resultados • A mostrar 0 - 100 (seguintes)

Texto:

Etiquetas:

contexto	a Universidade que é mais [—]	saudável!	: os professores ajudam-te
contexto	DE VISTA HISTÓRICO, E MUITO	AGRADÁVEL	É NÃO E QUASI
contexto	escritora J.M.	AVEL	, sobre a transição do
contexto	PARA MIM ESTA É UM OCASIÃO	IRREPETÍVEL	PARA APRENDER QUALQUER COISA DÁ
contexto	OU SEJA, É MUITO	PROVÁVEL	QUE QUALQUER NÃO GOSTERIA
contexto	PESSOA MUITO AMBICIOSA, SIMPÁTICA E	SOCIÁVEL	. SOU SEMPRE SORRIDENTE
contexto	. POR ISSO É UMA CIUDADE	SOPPORTÁVEL	QUE NÃO TRAZ MUITO
contexto	manter a forma num nível	aceitável	, subo as escadas
contexto	de termos espaços urbanos tranquilos e	acesíveis	a pé. Prefiro
contexto	, cultura e compras são mais	acesível	na cidade do
contexto	cidade esta é também	acesível	mais fácil e quando
contexto	É inegável que os hospitais facilmente	acesíveis	e os centros comerciais
contexto	ciudades é também mais	acesível	porque a maioria das
contexto	os outros comerciais) e melhor	acesível	mas no outro
contexto	. Acho que é uma experiência	aconselhável	para jovens. As
contexto	Espero muito que você passe momentos	agradáveis	em XXXXX, e
contexto	, a vida aqui é muito	agradável	. Estive a ver
contexto	tenho muito tempo para fazer [—] coisas	agradáveis	. Primeiro, já
contexto	minhas composições sempre estejam bonitos ou	agradáveis	de se ouvir,
contexto	fantástico – as pessoas são muito	agradáveis	, tenho bastante amigos
contexto	problemas, e são sempre muito	agradáveis	. A Baixa está
contexto	contaminação, as pessoas são muito	agradáveis	e próximas □ Na altura
contexto	gostam mais passear nas florestas	agradáveis	(como no
contexto	magra. Sou muito simpática.	agradável	e gosto da

9. Como pesquisar sequências de palavras de acordo com a ordem que ocupam na frase, combinando *Classe morfossintática / Lema / Forma do aluno*?

- ✍ Esta é uma forma de pesquisa avançada que terá que ser feita preenchendo *manualmente* o campo de pesquisa geral.
- ✍ A pesquisa de um dado segmento obedece à seguinte estrutura:

[etiqueta="classe morfossintática.*"]
[etiqueta="lema/forma do aluno"]

- ✍ Os códigos para as etiquetas são os seguintes:

pos = classe morfossintática
lemma = lema
form = forma do aluno
nform = forma corrigida

- ✍ As abreviaturas associadas às diferentes classes morfossintáticas podem ser consultados *aqui* (PDF).

- ✍ Exemplos de pesquisa:

1. **[pos="AQ.*"] [pos="N.*"]**
(adjetivo qualificativo seguido de nome)

Pesquisa no corpus

CQP Query: [pos="AQ.*"] [pos="N.*"] construção da pesquisa | ver | opções

1244 resultados • a mostrar 500 - 600 (seguintes)

Texto:

Etiquetas:

contexto	, como a	grande parte	dos outros estudantes,
contexto	outros países. [→] É um	grande pasatempo	para mim, e aunque
contexto	natureza. Um	grande paixão	da minha são entre
contexto	Mas eu [→] costumo tomar um	grande pequeno-almoço	. Gosto de cereais e
contexto	tem em comum é a	grande percentagem	de estudantes e uma universidade
contexto	eu detestar as vilas com	grande população	. Portanto veremos, se
contexto	4 meses. Era um	grande prazer	por mim. Estou longo
contexto	e isso da-me o	grande prazer	e satisfação. Será que
contexto	beber um copo é um	grande prazer	, que só podemos encontrar
contexto	compras, é a	grande quantidade	de merdes de
contexto	uma quarta [→] minha irmã,	grande quarta	com televisão e casa de
contexto	a Portugal) e um[→]	grande quarto	para jantar com muitas janelas
contexto	. As pessoas comem um	grande refeição	que consiste no peru
contexto	estudantes e também fica uma	grande residência	universitária. Todas as vezes
contexto	Apesar de desvantagens sinto um	grande sentimento	para o meu bairro,

2.

[pos="SP.*"] [lemma="casa"]

(nome "casa" antecedido de preposição)

Pesquisa no corpus

CQP Query: [pos="SP.*"] [lemma="casa"] construção da p

239 resultados • a mostrar 0 - 100 (seguintes)

Texto:

Etiquetas:

contexto	mais	em casa	neste sítio.
contexto	QUE PODE PREPARAR	EM CASA	COM A FARINHA DO
contexto	seus lugares são confortáveis.	Em casa	na Suíça tenho uma
contexto	família por os fim-de-semanas.	Em casa	, comemos almoso e jantar
contexto	horas e 30 minutos.	Em casa	vou fazer almoço Depois de
contexto	exame tenho muito actividade.	Em casa	e tambe fora dela
contexto	e o meu ávo.	Em casa	quando não faz nada eu
contexto	uma linha do telefone	a casa	e ninguem pode falar inglês
contexto	horas , e nós vão	a casa	depois de estuda. em
contexto	ouço a música e arrumo	a casa	, ás 9:55 ajudo
contexto	bonito e perto são lugares	com casas	velhas onde morava pessoas pobres
contexto	Verão está muito calor dentro	das casas	, assim que faço quase
contexto	causa dos meus trabalhos	de Casa	não são muitos, pois
contexto	minha mãe é a Dona	de Casa	. Com este trabalho o
contexto	de pipoca. Depois saio	de casa	, vou para doce vida

3.

[form="se"] [pos="V.*"]

(clítico+verbo)

Pesquisa no corpus

CQP Query: construção da pesquisa

376 resultados • a mostrar 0 - 100 (seguintes)

Texto:

Etiquetas:

contexto	iam cobertar a praia para	se refrescas-lhes	. A minha cidade
contexto	coisa para fazer e não	se aborecer	. No outro lado
contexto	de se beijar e de	se abraçar	. Frequentemente eles não são
contexto	burro mas com saberes,	se abre	com alegria). Os
contexto	semanas, o semestre já	se acabou	. Mas até então ainda
contexto	para saber como é que	se acabou	o meu relacionamento com o
contexto	, mas temos de perguntar	se aconteceu	alguma coisa mais interessante.
contexto	uma sorte para mim porque	se acordar	tarde, posso correr e
contexto	, tenho de dizer que	se acrescenta	ao Som linguístico o
contexto	o mais importante cá é	se adaptar	à cultura portuguesa que
contexto	campo não são [-] perigosas,	se analisamos	as desvantagens da cidade
contexto	casa. Por exemplo,	se ando	cinco minutos da minha
contexto	cidade. Mas assim nunca	se apanha	a vida completa nas
contexto	caso de que	se apanhe	algum meio de transporte)
contexto	que os seres humanos não	se aprisionem	em as suas separações de
contexto	, na cidade não	se aproveita	esta vantagem do campo
contexto	encontrar as tradições antigas que	se arrastam	no decurso do

📖 Nesta pesquisa, observa-se que a **forma do aluno** corresponde, não só ao pronome pessoal, mas também à conjunção subordinativa.

📖 É possível introduzir uma restrição à **forma do aluno** para obter apenas resultados para a estrutura pretendida (clítico+verbo).

Pesquisa no corpus

CQP Query: construção da

209 resultados • A mostrar 0 - 100 (seguintes)

Texto:

Etiquetas:

contexto	coisa para fazer e não	se aborecer	. No outro lado
contexto	de se beijar e de	se abraçar	. Frequentemente eles não são
contexto	burro mas com saberes,	se abre	com alegria). Os
contexto	semanas, o semestre já	se acabou	. Mas até então ainda
contexto	para saber como é que	se acabou	o meu relacionamento com o
contexto	, tenho de dizer que	se acrescenta	ao Som linguístico o
contexto	o mais importante cá é	se adaptar	à cultura portuguesa que
contexto	cidade. Mas assim nunca	se apanha	a vida completa nas
contexto	caso de que	se apanhe	algum meio de transporte)
contexto	, na cidade não	se aproveita	esta vantagem do campo
contexto	encontrar as tradições antigas que	se arrastam	no decurso do
contexto	campo pode ser maçador quando	se atinge	a idade da adolescência
contexto	culinária especial, que naturalmente	se baseia	na presença dos
contexto	outro, de	se beijar	e de se abraçar.
contexto	conseguimos encontrar uma tasca onde	se canta	fado. No verão
contexto	. Tenho uma irmã que	se chama	Christine e tem 31 anos
contexto	uma festa muito interessante que	se chama	"A Queima das Fitas",

4.

[pos="V.*"] [lemma="se"]

(verbo+clítico)

Pesquisa no corpus

CQP Query: [pos="V.*"] [lemma="se"] construção

327 resultados • A mostrar 0 - 100 (seguintes)

Texto: Etiquetas:

contexto	largíssima e muito interessante –	nota-se	isso na arquitectura,
contexto	. O meu outro prazer	chama-se	cozinha. [→] Agora
contexto	a minha vida em Lyon	tornou-se	mais simples e já não
contexto	Portugal, a minha cidade	chama-se	Coimbra e é a terceira
contexto	, mas na noite	torna-se	perigoso e com muito barulho
contexto	vida no campo	limita-se	a um mês por ano
contexto	tenho um irmão, que	chama-se	XXXXX e é estudante de
contexto	os que moram alí.	Pode-se	dizer que a privacy não
contexto	, de manhã	ouve-se	muito barulho: passam por
contexto	: acostumo fazer jantares onde	reunir-se	tudos e saber como vão
contexto	de planos, embora possam	fazer-se	em outros lugares como de
contexto	ferias. A minha irmã	chama-se	XXXXX e tem 21 anos
contexto	eu gosto de ele.	Chama-se	XXXXX. Os meus pais
contexto	. Eu não gosto de	levantar-se	muito cedo e jogar futebol
contexto	prazer. [→] Em esta cidade	pode-se	visitar a torre, a
contexto	vezes ruda, de	comportar-se	deles. Trabalhar para

10. Como devem ser pesquisados os sinais de pontuação?

- ✍ Embora os sinais de pontuação tenham sido todos lematizados, a sua pesquisa deve ser feita por *Classe morfossintática > Pontuação > Tipo*.

Pesquisa no corpus

CQP Query: [construção da pesquisa](#) | [ver](#) | [opções](#)

Construção de etiquetas: Classe morfossintática

POS principal:

Tipo:

Posição:

Inserir:

aspas
parêntesis
hífen
barra
ponto interrogação
chaveta
ponto final
parêntesis retos
reticências
percentagem
ponto e vírgula
outros

Construção de etiquetas

Pesquisa do documento

Nacionalidade:

Língua materna:

Proficiência:

Fase de recolha:

Estímulo:

Adicionar token

| [ajuda](#)

Pesquisa no corpus

CQP Query: [construção da pesquisa](#) | [ver](#) | [opções](#)

282 resultados • A mostrar 0 - 100 (seguintes)

Texto:

Etiquetas:

contexto | A vida no campo : | Eu nasci numa
contexto um país com muitos rostos diferentes : O nível cultural e
contexto ser necessariamente muito pessoal e subjectivo : Minha família mora no
contexto minha família (excepções : um castelo velho e
contexto a primeira vez, eu pensei : "Aqui há mais
contexto A primeira e mais importante é : já sou tial
contexto pessoas usaram os animais para viajar : camelos, cavalos □ Depois
contexto quase todos os meios de transporte : a bicicleta, o
contexto relação muito forte com a natureza : o mar, o
contexto viajem a cavalo, um longe : sair de casa a
contexto , mas é um bocadinho diferente : não há a mesma
contexto a um mês por ano : quando venho de férias
contexto de manhã ouve-se muito barulho : passam por aí muitas
contexto família e dos amigos : acostumo fazer jantares onde reunir-se
contexto mais queridas para curtir estos planos : qué melhor que escolher
contexto a Paris com a minha mãe : foi interessante e consegui

11. Existe alguma forma de pesquisar palavras não coincidentes com o português europeu no que se refere à ortografia e à acentuação gráfica? *Fácil ou difícil?*

- ✍ A única forma de pesquisar palavras não coincidentes com o português europeu é construir, de acordo com o âmbito do estudo, uma pesquisa por **Lema**.
- ✍ Os resultados são todas as formas, da(s) palavra(s) em questão, coincidentes e não coincidentes com o português europeu.
- ✍ Refinar a pesquisa de acordo com os objetivos definidos.

Pesquisa no corpus

CQP Query: [construção da pesquisa](#) | [ver](#) | [opções](#)

Construção de pesquisa

Pesquisa do texto

Forma do aluno

Forma corrigida

Classe morfosintática

Lema

[ajuda](#)

Pesquisa do documento

Nacionalidade

Língua materna

Proficiência

Fase de recolha

Estímulo

Pesquisa no corpus

CQP Query: [construção da pesquisa](#) | [ver](#) | [opções](#)

66 resultados

Texto:

Etiquetas:

contexto	bastante boas – não são muito	facéis	compreender, mas acho
contexto	Porém, a vida não anda	facil	desde que estou a
contexto	em Coimbra eu posso aprender mais	facil	o livre para estrangeiros
contexto	quero trocar extrovertido e faço mais	facil	amigos. Outras pessoas
contexto	a falar em português é mais	facil	agora comparado do
contexto	>viagem da Europa não era	facil	, mas era muito
contexto	mundo, com transporte publico	facil	e frequente e cada
contexto	tem aeroportos por isso é muito	facil	e conveniente para viajar
contexto	de casa. Também é mais	facil	encontrar um emprego numa
contexto	barulho, anda muito	facil	morar no velho
contexto	Nadar no Atlantico não e	facil	mais eu prefiro o [---]
contexto	posso dizer que e muito mais	facil	ler do que
contexto	Agora a nossa vida não é	facil	mas temos muitas pessoas
contexto	para divertir durante fim de semana. É	facil	encontrar um bom emprego
contexto	quando fui para Lisboa foi muito	facil	ir para ali de
contexto	em Lisboa. Esta cidade é	facil	de visitar por os
contexto	minha casa. Não e muito	facil	de trabalhar efectivo por

Pesquisa no corpus

CQP Query: [lemma = "difícil"] within text

[Pesquisar](#) construção da pesquisa | ver | opções

106 resultados • A mostrar 0 - 100 (seguintes)

Texto: [Transcrição](#) | [Forma do aluno](#) | [Forma corrigida](#)

Etiquetas: [Classe morfosintática](#) | [Lema](#)

contexto	VIVER NA CIDADE [] PODE SER	DIFÍCIL	PARA UMA PESSÓA COME
contexto	AS RELAÇÕES PESSOAIS VAI SER MAIS	DIFÍCIL	DAS SIMPLES E
contexto	oceáraria. É muito	difícil	nadar aqui. Mas
contexto	a nossa história sempre era muito	difícil	– a Polónia ficava na
contexto	para Coimbra mas, porque é	difícil	encontrar lugares de estacionamento
contexto	viver no campo, é	difícil	ir à hospital
contexto	os quartos. Para mim é	difícil	habituar-se com transporte
contexto	início, as aulas foram	difíceis	. Com certeza,
contexto	e estudo dele é muito	difícil	e precisa energia para
contexto	argentinos porque é realmente	difícil	até para mim que
contexto	Cracovia. Agora tenho o tempo	difícil	, porque escreve os
contexto	ele esta em uma idade muito	difícil	. mas eu gosto
contexto	, entons pela noite é	difícil	ver bem. ! No
contexto	, mas próximo mese estive muito	difícil	. Eu estudou muitíssimo
contexto	Junho, porque houve tempo muito	difícil	e depressando . Em
contexto	sozinho. No principio muito	difícil	, mas agora estou
contexto	são o estudo, porque é	difícil	estudar em Erasmus com

12. A pesquisa de dados de natureza pragmática é possível? *Adeus e mil beijinhos.*

- ✎ Não existe uma etiqueta pesquisável de natureza pragmática, mas algumas pesquisas podem ser úteis neste domínio por permitirem observar diferentes tipologias textuais, formas de tratamento, registo formal/informal e até atos ilocutórios em diferentes contextos. Seguem alguns exemplos de pesquisa.

1. por *Estímulo*

- ✎ **Construção de pesquisa > Estímulo > 6.1B** (carta informal)

Pesquisa no corpus

CQP Query: [construção da pesquisa](#) | [ver](#) | [opções](#)

Construção de pesquisa

Pesquisa do texto

Forma do aluno:

Forma corrigida:

Classe morfosintática:

Lema:

[ajuda](#)

Pesquisa do documento

Nacionalidade:

Língua materna:

Proficiência:

Fase de recolha:

Estímulo:

- 1.1A
- 1.1A
- 3.1J
- 5.2L
- 5.2L
- 5.2M
- 6.1B**
- 6.9.3Q
- 7.5.3S
- 7.7.3T

Pesquisa no corpus

CQP Query: [construção da pesquisa](#) | [ver](#) | [opções](#)

133 resultados • A mostrar 0 - 100 * (seguintes)

ID	Nacionalidade	Língua materna	Proficiência	Fase de recolha	Estímulo
espanhol.c1.13.6.1b.xml	Colombiana	Espanhol	C1	Fase 1	6.1B
romeno.a1.14.6.1.b.xml	Romena	Romeno	A1	Fase 1	6.1B
espanhol.c1.18.6.1b.xml	Espanhola	Espanhol	C1	Fase 1	6.1B
atalao.b2.75.6.1b.xml	Espanhola	Catalão	B2	Fase 2	6.1B
italiano.b1.145.6.1b.xml	Italiana	Italiano	B1	Fase 1	6.1B
polaco.a1.23.6.1b.xml	Polaca	Polaco	A1	Fase 1	6.1B
italiano.b2.38.6.1b.xml	Italiana	Italiano	B2	Fase 1	6.1B
polaco.b2.13.6.1b.xml	Polaca	Polaco	B2	Fase 1	6.1B
ingles.b1.60.6.1b.xml	Galesa	Inglês	B1	Fase 1	6.1B
neerlandes.b1.115.6.1b.xml	Holandesa	Neerlandês	B1	Fase 1	6.1B
ingles.b2.79.6.1b.xml	Britânica	Inglês	B2	Fase 2	6.1B
galego.b1.140.6.1b.xml	Espanhola	Galego	B1	Fase 1	6.1B
italiano.a1.42.6.1b.xml	Italiana	Italiano	A1	Fase 1	6.1B
turco.a1.24.6.1b.xml	Turca	Turco	A1	Fase 1	6.1B
italiano.a1.18.6.1b.xml	Italiana	Italiano	A1	Fase 1	6.1B
espanhol.b2.77.6.1b.xml	Espanhola	Espanhol	B2	Fase 2	6.1B
ingles.b1.129.6.1b.xml	Inglesa	Inglês	B1	Fase 1	6.1B
lituano.a1.16.6.1b.xml	Lituana	Lituano	A1	Fase 1	6.1B
chines.a1.10.6.1b.xml	Chinesa	Chinês	A1	Fase 1	6.1B
grego.a1.17.6.1b.xml	Grega	Grego	A1	Fase 1	6.1B
italiano.b2.60.6.1b.xml	Italiana	Italiano	B2	Fase 2	6.1B

- Combinar esta opção com o campo **Lema**, preenchido com uma forma de despedida (por exemplo, “beijo”).

Pesquisa no corpus

CQP Query: [construção da pesquisa](#) | [ver](#) | [opções](#)

Construção de pesquisa

Pesquisa do texto		Pesquisa do documento	
Forma do aluno	<input type="text" value="igual a"/>	Nacionalidade	<input type="text" value="[selecionar]"/>
Forma corrigida	<input type="text" value="igual a"/>	Língua materna	<input type="text" value="[selecionar]"/>
Classe morfosintática	<input type="text" value="construção de etiquetas"/>	Proficiência	<input type="text" value="[selecionar]"/>
Lema	<input type="text" value="igual a"/>	Fase de recolha	<input type="text" value="[selecionar]"/>
	<input type="text" value="beijo"/>	Estímulo	<input type="text" value="6.1B"/>

[cancelar](#) | [ajuda](#)

Pesquisa no corpus

CQP Query: [construção da pesquisa](#) | [ver](#) | [opções](#)

85 resultados

Texto:

Etiquetas:

contexto	noticias tuas, Forte abraço, Beijinhos ,
contexto	vou a dizer mais. Beijinhos e até breve
contexto	quando eu voltar para Inglaterra. Beijinhos ,
contexto	este tudo bem contigo agora. Beijinhos
contexto	Europa em fim de Agosto. Beijinhos ,
contexto	não acabamos com as historias? Beijinhos ,
contexto	- meus filhos ainda não comeram □ Beijinhos ,
contexto	ti e Feliz Natal antecipadamente. Beijinhos e abraços
contexto	ti para me contar tudo. Beijinhos , a tua amiga
contexto	a tua namorada? Ciao! Beijinhos ! a tua amiga
contexto	. Até o Natal Beijinhos [---] XXXXX (Vou chegar no
contexto	Não ezquesas de vir a Coimbra
contexto	Natal na Áustria! Beijinhos e abraços!
contexto	Volto para Bucareste em Julho [---] Beijinhos .
contexto	net. Adoro-te muito. Beijinhos minha irmã! A
contexto	ao vivo também! Beijinhos e abraços
contexto	Estou ansiosa a falar contigo. Beijinhos para ti e os

- Obter a frequência dos resultados por **nacionalidade**.

Expressão de busca: Lema = beijo
 Estimulo = 6.1B
 Agrupamento de pesquisas: Nacionalidade

Gráfico: Tabela | Contagem: Contagem | Guardar: [selecionar]

Grupo	Contagem
Turca	1
Sueca	1
Romena	4
Portuguesa	1
Polaca	8
Lituana	1
Letã	2
Italiana	10
Iraniana	1
Inglesa/ Portuguesa	1
Inglesa	8
Húngara	1
Grega	2
Galesa	1
Francesa/Portuguesa	2
Francesa	2
Espanhola	11
Eslovaca	1
Colombiana	1
Chinesa	6
Checa	6
Canadiana	1
Búlgara	1
Britânica	1
Austriaca	2
Alemã	9

2. por **Verbos** que denotam atos de fala.

Pesquisa no corpus

CQP Query: [lemma = "prometer"] within text Pesquisar [construção da pesquisa](#) | [ver](#) | [opções](#)

2 resultados

Texto: Transcrição Forma do aluno

Etiquetas: Classe morfosintática Lema

contexto te ter dito nada antes! **Prometo** não ser tão desleixada no
 contexto amiga! Quando visitares, eu **promisso** que mostrar-te-ei as

[Descarregar resultados](#) - [Memorizar expressão de busca](#)

Opções de frequência

Colocação por: [selecionar] | Tamanho do contexto: 1 | Direção: esquerda e direita Submeter

Frequência por: [selecionar]

3. por **Pronomes pessoais e formas de tratamento**

Pesquisa no corpus

CQP Query: [lemma = "você"] within text construção da pesquisa |

33 resultados

Texto:

Etiquetas:

contexto	de ser um dos	você	.
contexto	domingo, seja	você	religioso ou não, participar
contexto	e horas a brincar com	você	. Ele era tão pequeno
contexto	nós escrevimos. Como vai	você	desde a nossa última carta
contexto	carta? Espero muito que	você	passse momentos agradáveis em XXXXX
contexto	"Oh, boal,	você	tem mesmo muita sorte
contexto	foram bastante corridos. Como	você	sabe, [:-] eu estive
contexto	em brasileiro. Minha querida,	você	não tem noção de quan
contexto	que esqueci completamente de perguntar como	você	está. O quê
contexto	outra vantagem que posso deixar	você	persuadir a viver na
contexto	Qualquer coisa é preparada para	você	perto de casa. E
contexto	oferecem uma vantagem importante—	você	pode fazer outras coisas ao
contexto	limpio. Então já será como	você	: um desempregado mais
contexto	mais com dicionário. Quando	você	pergunta à alguém sobre
contexto	para ter sucesso. Qual	você	ajudar o teu familia nos
contexto	Região. Aqui	você	pode sentir a brisa enquanto
contexto	guerra civil Para	você	, visitar tudo aquilo não

4. por *Sinais de pontuação*.

Pesquisa no corpus

CQP Query: construção da pesquisa | ver | opções

Construção de etiquetas: Classe morfosintática

POS principal:

Tipo:

Posição:

Constru

Pesquis

Classe m

Pesquisa do documento

Nacionalidade:

Lingua materna:

Proficiência:

Fase de recolha:

Estimulo:

contexto	juntos. O que tu achas	?	Nos últimos dias,
contexto	Querida XXXXX, Como estás	?	Lembraste o dia do
contexto	o dia do teu aniversário	?	Lembraste que nós
contexto	? Lembraste que nós fazemos	?	Fomos ao restaurante
contexto	hábitos significativos da minha cultura	?	Os alemães chegam (
contexto	É tudo que eu escrevi verdade	?	Nunca se sabe □ (
contexto	gostava quero fazer nestes momentos	?	É uma viagem,
contexto	com pensamentos positivos, lembras-te	?	Bons momentos que passámos
contexto	"jantarada" curso	?	é inadmissível ficar tanto
contexto	Querida XXXXX, como estás	?	Já passaram quasi dois
contexto	XXXXX, Então, como estas	?	Tudo bem para ti
contexto	para ti e para tua família	?	Por aqui estou muito
contexto	Tavira tu foi também passado	?	Que pena eu vou
contexto	Sou o XXXXX, como estas	?	Há muito tempo que
contexto	nós não vemos, não achas	?	Agora tu não te
contexto	quando fomos [→] [→] XXXXX em França	?	Eu gostei muito daquelas
contexto	de todos os concertos que vimos	?	Fantastico! Agora que
contexto	? Fantastico! Agora que fazes	?	Não sei se tu [→]
contexto	Cara XXXXX, Tudo bem contigo	?	Não nos vemos há
contexto	. O que andas a fazer	?	Este ano acabas a
contexto	acabas a escola, não é	?	Eu estou bem.
contexto	como uma vez nos perdemos	?	Tivemos tanto medo,
contexto	E, tudo isto para que	?	Porque gosto destas
contexto	? Porque gosto destas coisas	?	Só há uma resposta
contexto	porque se não, para que viver	?	A conclusão final e
contexto	XXXXX. Como estás minha amiga	?	Tenho muitas saudade para
contexto	E como está tua família	?	Eu gosto muito de

Pesquisa no corpus

CQP Query: [pos = "Fs.*"] within text

[Pesquisar](#) [construção da pesquisa](#) | [ver](#) | [opções](#)

9 resultados

Texto: [Transcrição](#) [Forma do aluno](#)

Etiquetas: [Classe morfosintática](#) [Lema](#)

contexto	frio ou chova no exterior	...	Passar o tempo a
contexto	falar de tudo e de nada	...	Também dá jeito ir
contexto	onde. Tantos momentos vivimos juntas	...	mas de todos,
contexto	mente durante este tempo	...	Falando de viagem,
contexto	serra, falar com meus amigos Agora, ciclismo
contexto	Repúblicas têm tudo para fazer isto	...	têm espaço.
contexto	já és um homem trabalhador?	...	Acho que isso nunca
contexto	Que irreverentes que éramos!	...	Tu tiveste sempre uma
contexto	gostam de comer pasta, pão etc.	..	à

5. por *Interjeição*.

Pesquisa no corpus

CQP Query: [pos = "I.*"] within text

[Pesquisar](#) [construção da pesquisa](#) | [ver](#) | [opções](#)

96 resultados

Texto: [Transcrição](#) [Forma do aluno](#) [Forma corrigida](#)

Etiquetas: [Classe morfosintática](#) [Lema](#)

contexto	tu vais venir visitar-me.	Adeus	e mil beijinhos.
contexto	paraia. E tu?	Adeus	Um beijinho
contexto	encontramos o caminho de volta.	Ah	, que pena que
contexto	vida cá e até logol	Ah	, esqueci-me uma
contexto	me sentia a vontade.	Ah	, mesmo não quero
contexto	agora pertence ao passado.	Ah	, nossa vida estudantil
contexto	toda a gente a dançar.	Ah	! Dançar. Também
contexto	toros, flamenco e paella.	Ah	! e a praia
contexto	sempre, temos de pagarl	Ah	, outra coisa que
contexto	dias juntos, que chatice!	Ah	, ah antes na
contexto	otos desse teste?!	Ahah	sèm dúvidas un dos
contexto	seis da manhãl	Ai	, tenho que ir
contexto	está na minha cabeça.	Ai	, amiga, penso
contexto	devo chorar mas é incontrolável)	Bem	□ Sei que toda a
contexto	pensam aqui □ Estou a brincar!	Bom	Amigo espero que possas
contexto	natal tenho poucas feiras.	Caralho	!! Eu tenho
contexto	e que a tua namorada?	Ciao	! Beijinhos! a

13. Como obter a frequência de determinadas ocorrências em função de diferentes critérios? *É fácil!*

- 📄 Após a exibição dos resultados de uma pesquisa, percorrer os resultados até ao final da página onde se encontram os campos relativos a **Opções de frequência**.
- 📄 Ao seleccionar a opção pretendida, serão apresentados os resultados em tabela.
- 📄 **Nota:** Presentemente, a opção **Custom distribution** está a ser atualizada pelo que ainda não é possível utilizá-la.

contexto	Geralmente gosto de meu bairro porque	e fácil	andar todos os dias
contexto	A viagem da Europa não	era fácil	, mas era muito
contexto	eu tinha entendido que de Erasmus	era fácil	engatar com as meninas
contexto	são muito simpáticos, por isso	foi fácil	para fazer amizades com
contexto	e de Billy Joel porque não	são fácil	para aprender. Também
contexto	nossa para divertir durante fim de semana.	É fácil	encontrar um bom emprego
contexto	Há muitas coisas para fazer.	É fácil	encontrar pessoas simpáticas e
contexto	segundo a cozinha de Mejiço.	É fácil	encontrar novos amigos quando sabe-se
contexto	tempo com a minha namorada.	É fácil	gostar de cada actividade [→]
contexto	. Agora a nossa vida não	é fácil	mas temos muitas pessoas
contexto	chegar em Lisboa. Esta cidade	é fácil	de visitar por os
contexto	simple. [→] ir para Pisa	é facilimo	!! Ali ha
contexto	inglês ou em português e não	é fácil	para mim. Eu
contexto	que é um bairro [→] onde não	é fácil	viver porque há muita
contexto	a nossa situação política: não	é fácil	. [→] Há três [→] línguas
contexto	barulho e as vezes não	é fácil	dormir. Mas viver na
contexto	você perto de casa. E	é fácil	fazer vários tipos de
contexto	há poluição, vida	é fácil	mas no ponto
contexto	seu próprio meio de transporte,	é fácil	ficar isolado e pode
contexto	É por isso que para mim	é fácil	acostumar-me à

Descarregar resultados [seleccionar] expressão de busca

Opções de frequência: Estímulo

Colocação por: Texto | Número do contexto: 1 | Direção: esquerda e direita | Submeter

Frequência por: [seleccionar]

Distribuição no corpus

Expressão de busca: Lema = **fácil**

Agrupamento de pesquisas: Estímulo

Gráfico: Tabela | Contagem: Contagem | Guardar: [seleccionar]

Grupo	Contagem
77.3T	3
75.3S	3
69.3Q	5
6.1B	3
52.2L	1
50.2L	1
33.1J	3
1.1A	2

[Ajuda](#) • [URL direto](#)

14. Em que formatos podem ser descarregados os textos?

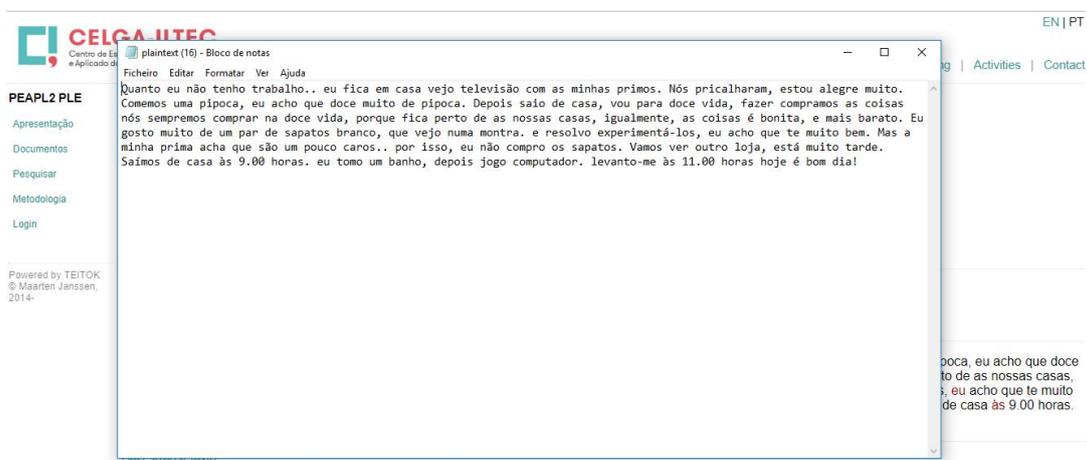
- 📄 As produções escritas podem ser descarregadas em dois formatos diferentes:
 - i. em formato **txt**.
- 📄 Ao consultar um texto, escolher a **Opção de representação** que pretende obter (*Transcrição, Forma do aluno, Forma corrigida, Cores, Classe morfossintática ou Lema*).
- 📄 Seguidamente, clicar em **Descarregar texto**, no final da página.

Opções de representação

Texto: Transcrição Forma do aluno Forma corrigida - Mostrar: Cores - Etiquetas: Classe morfossintática Lema

Quando eu não tenho trabalho.. eu fica em casa vejo televisão com as minhas primos. Nós pricalharam, estou alegre muito. Comemos uma pipoca, eu acho que doce muito de pipoca. Depois saio de casa, vou para doce vida, fazer compramos as coisas nós sempremos comprar na doce vida, porque fica perto de as nossas casas, igualmente, as coisas é bonita, e mais barato. Eu gosto muito de um par de sapatos branco, que vejo numa montra. e resolvo experimentá-los, eu acho que te muito bem. Mas a minha prima acha que são um pouco caros.. por isso, eu não compro os sapatos. Vamos ver outro loja, está muito tarde. Saímos de casa às 9.00 horas. eu tomo um banho, depois jogo computador. levanto-me às 11.00 horas hoje é bom dia!

Descarregar texto



ii. em formato **Word**.

- 📄 Neste caso, apenas obtém a transcrição do texto do aluno (respeitando as convenções de transcrição), sem anotações linguísticas, visto este formato ter sido disponibilizado na fase inicial do projeto de *Recolha do Corpora de PL2*. Estes ficheiros, organizados por fase, língua materna e nível de proficiência dos aprendentes, podem ser obtidos [aqui](#).

15. Como aceder ao perfil dos informantes?

 Existem duas formas de aceder ao perfil dos informantes:

i. consultar, individualmente, o cabeçalho (reduzido ou expandido) de cada uma das produções escritas;

alemao.b2.66.69.3q	
alemao.b2.66.69.3q	
Língua materna	Alemão
Género	F
Nacionalidade	Alemã
QECRL	B2
▶ mais dados	

alemao.b2.66.69.3q	
alemao.b2.66.69.3q	
Código do texto	ALEMÃO ER. B2.66
Estímulo	69.3Q
Nº de informantes	1
Fase de recolha	Fase 2
Nº médio de palavras	423
QECRL	B2
Student	
Data de nascimento	1989.02.02
Ano de início de estudo do português	2009
Fala português fora do contexto escolar?	Sim, todos os meus colegas de casa e o meu namorado são portugueses
Género	F
Língua de escolarização	Alemão
Língua materna	Alemão
País em que nasceu	Alemanha
Nacionalidade	Alemã
Países em que já viveu	*França/ 10 meses; Portugal/ 17 meses*
PT Proficiency	
Produção escrita	B2
Compreensão escrita	C1
Produção oral	B2
Interação oral	B2
Compreensão oral	B2
Other Foreign Language(s)	
Outras línguas não maternas?	Alemão/ Inglês/ Francês
Língua estrangeira em que tem maior proficiência	Inglês
Produção escrita	C1
Compreensão escrita	C2
Produção oral	C1
Interação oral	C1
Compreensão oral	C1

ii. descarregar o ficheiro com os dados dos informantes, em formato *Excel*, que se encontra disponível na página inicial do projeto, [aqui](#) (na secção **Informantes**).

16. Como descarregar a listagem de resultados?

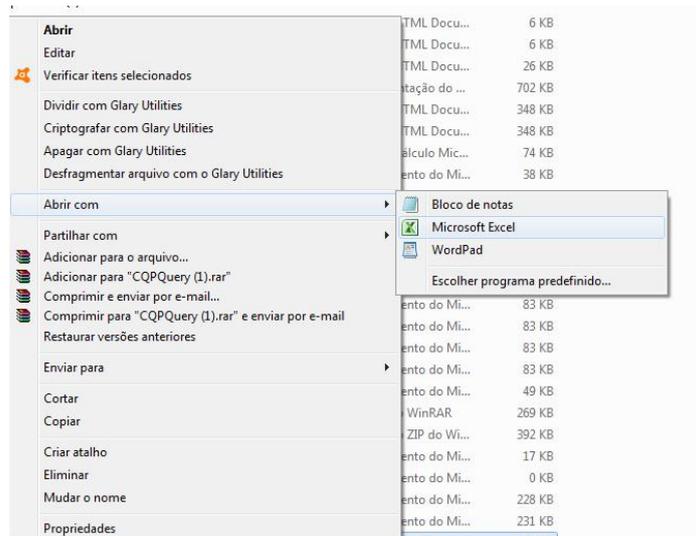
Quando se realiza uma pesquisa, é apresentada uma listagem com os resultados obtidos.

- 📄 Para além de poder consultar cada uma das produções, é possível **Descarregar resultados**, no final da página.

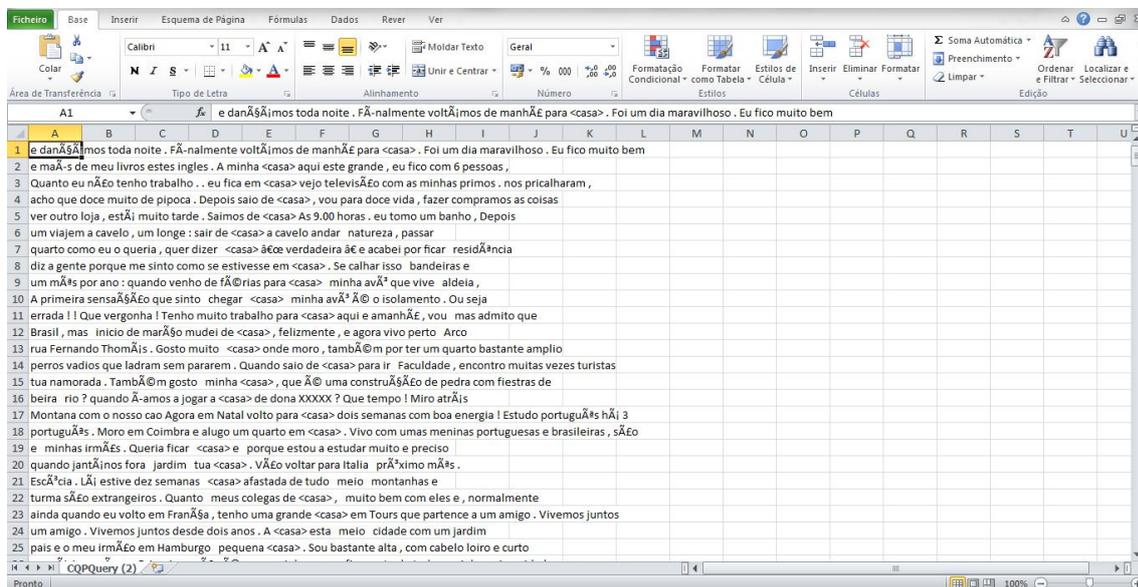
contexto	Ensinei as amigas portuguesas da	casa	e elas acharam muita
contexto	estás? Como é que em	casa	? Por aqui estou
contexto	. Quando eu mudei da	casa	dos meus pais
contexto	Os meus pais moram numa	casa	grande, em Lisboa
contexto	grande, em Lisboa. A	casa	tem 4 quartos,
contexto	Não gostava muito de voltar par	casa	sozinha à noite
contexto	Continente e volto para	casa	de autocarro
contexto	cidade. A minha	casa	fica muito perto da
contexto	arrumar as coisas. En tudo	casa	, manda um grande
contexto	, pode fazer os trabalhos de	casa	. Algumas pessoas lêem
contexto	para fazer os meus trabalhos de	casa	. Depois as aulas
contexto	as aulas, eu volto para	casa	para surf o Internete
contexto	para fazer os trabalhos na	casa	
contexto	os dias não saio da	casa	, e a única
contexto	há menos e para voltar a	casa	à noite as
contexto	e espero que estar na	casa	porque vou visitar os

[Descarregar resultados](#) - Memorizar expressão de busca

- 📄 Descarregue os resultados e seleccione **Abrir com** e seleccione a opção pretendida.
- 📄 Sugerimos o **Excel**, pois permite fazer várias operações ao nível do tratamento de dados .



Cada página de resultados apresenta 100 ocorrências de cada vez. Ao descarregar os resultados, basta fazê-lo uma vez para guardar, neste exemplo, as 481 ocorrências da palavra “casa”, não sendo necessário fazê-lo página a página.



Nota: A palavra pesquisada, surge identificada entre < parênteses angulares >.

17. Como descarregar os dados de frequência de uma ocorrência? *É fácil!*

- 📄 Percorrer os resultados até ao final da página onde se encontram os campos relativos a **Opções de frequência**.
- 📄 Ao seleccionar a opção pretendida, serão apresentados os resultados em tabela.

contexto Geralmente gosto de meu bairro porque **e fácil** andar todos os dias
contexto A viagem da Europa não **era fácil** , mas era muito
contexto eu tinha entendido que de Erasmus **era fácil** engatar com as meninas
contexto são muito simpáticos, por isso **foi fácil** para fazer amizades com
contexto e de Billy Joel porque não **são fácil** para aprender. Também
contexto nossa para divertir durante fim de semana. **É fácil** encontrar um bom emprego
contexto Há muitas coisas para fazer. **É fácil** encontrar pessoas simpáticas e
contexto segundo a cozinha de Mejico. **É fácil** encontrar novos amigos quando sabe-se
contexto tempo com a minha namorada. **É fácil** gostar de cada actividade [↔]
contexto . | Agora a nossa vida não **é fácil** mas temos muitas pessoas
contexto chegar em Lisboa. Esta cidade **é fácil** de visitar por os
contexto simple. [↔] ir pará Pisa **é facilimo** !! | Ali ha
contexto inglês ou em portugues e não **é fácil** para mim. Eu
contexto que é um bairro [↔] onde não **é fácil** viver porquê há muita
contexto a nossa situação política: não **é fácil** . [↔] Há três [↔] línguas
contexto barulho e as vezes não **é fácil** dormir. Mas viver na
contexto você perto de casa. E **é fácil** fazer vários tipos de
contexto há poluição. vida **é fácil** mas no ponto
contexto seu próprio meio de transporte, **é fácil** ficar isolado e pode
contexto É por isso que para mim **é fácil** acostumar-me à

Descarregar res: [seleccionar] Nacionalidade expressão de busca
Lingua materna
Proficiência
Opções de fr: Fase de recolha
Estímulo

Colocação por: Texto nº do contexto: 1 | Direção: esquerda e direita | Submeter
Custom distribution
Frequência por: [seleccionar]

Distribuição no corpus

Expressão de busca: Lema = **fácil**
Agrupamento de pesquisas: Estímulo

Gráfico: Tabela | Contagem: Contagem | Guardar: [seleccionar]

Grupo	Contagem
77.3T	3
75.3S	3
69.3Q	5
6.1B	3
52.2L	1
50.2L	1
33.1J	3
1.1A	2

[Ajuda](#) • [URL direto](#)

De seguida, em **Guardar**, seleccionar a opção **CSV**, para transferir os dados .

Distribuição no corpus

Expressão de busca: Lema = **fácil**
Agrupamento de pesquisas: Estímulo

Gráfico: Tabela | Contagem: Contagem | Guardar: [selecionar]

Grupo	Contagem
77.3T	3
75.3S	3
69.3Q	5
6.1B	3
52.2L	1
50.2L	1
33.1J	3
1.1A	2

Ajuda • URL direto

Download data as Comma-Separated Values

Após a transferência, clicar em **Abrir** .

PEAPL2 PLE

Apresentação
Documentos
Pesquisar
Metodologia
Login

Powered by TEITOK
© Maarten Janssen,
2014-

Distribuição no corpus

Expressão de busca: Lema = **fácil**
Agrupamento de pesquisas: Estímulo

Gráfico: Tabela | Contagem: Contagem | Guardar: CSV

Grupo	Contagem
77.3T	3
75.3S	3
69.3Q	5
6.1B	3
52.2L	1
50.2L	1
33.1J	3
1.1A	2

Ajuda

transferir

Abrir
Abrir sempre ficheiros deste tipo
Mostrar numa pasta
Cancelar

Finalmente, seleccionar o **Excel**, na caixa de diálogo, para abrir o ficheiro .

Distribuição no corpus

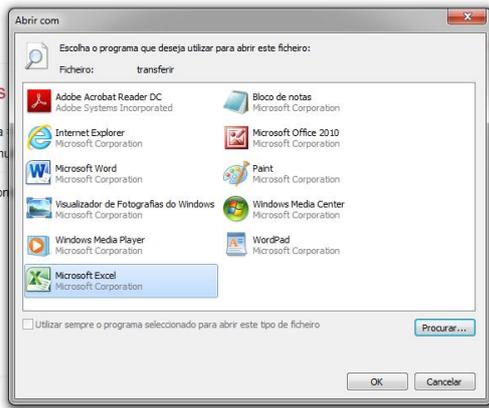
Expressão de busca: Lema

Agrupamento de pesquisas: Estimul

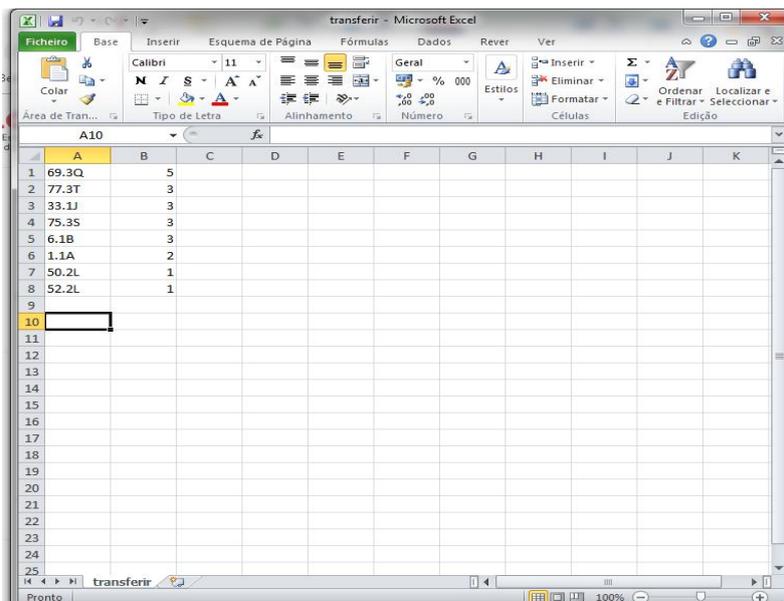
Gráfico: Tabela

Grupo	Contagem
77.3T	3
75.3S	3
69.3Q	5
6.1B	3
52.2L	1
50.2L	1
33.1J	3
1.1A	2

Ajuda • URL direto



📄 Guardar o ficheiro com o nome pretendido .



18. Como guardar e comparar expressões de pesquisa feitas anteriormente? *É simples... e fácil!*

- ☞ Percorrer a listagem de resultados obtidos e, no final da página, clicar em **Memorizar expressão de busca**.

Pesquisa no corpus

CQP Query: [construção da pesquisa](#) | [ver](#) | [opções](#)

11 resultados

Texto:

Etiquetas:

contexto	vida em Lyon tornou-se mais	simples	e já não haveram
contexto	onde se podem apreciar os coisas	simples	do dia à
contexto	; gosto mais das comidas	simples	. Eu sou italiana
contexto	, a interação não era tanto	simples	. Nunca tinha pensado
contexto	começo a apreciar voltar as coisas	simples	da vida,
contexto	idades muito tempo, pelo	simples	facto de eu detestar
contexto	qualquer prato, do mais	simples	até o mais complicado
contexto	cidade, e as razões são	simples	: é um compromisso
contexto	naturalmente, não é assim tão	simples	porque nesta o problem
contexto	as patologias básicas de forma muito	simples	para nós entendermos. Confesso-te
contexto	com eles nos tivemos uma vida	simples	. Em portuga eu

[Descarregar resultados](#) - [Memorizar expressão de busca](#)

- ☞ Repetir o processo a cada nova pesquisa.
- ☞ Quando terminar, no final da página, clicar em **Expressões CQL guardadas**.

contexto	tipo de interlocutor, é mais	fácil	de ometter uma informação
contexto	filhos. Embora seja mais	fácil	viver na cidade
contexto	muito simpáticos, por isso foi	fácil	para fazer amizades com
contexto	não é uma coisa evidente o	fácil	porque sempre há uma
contexto	a cozinha de Mejico. É	fácil	encontrar novos amigos quando sabe-se
contexto	com a minha namorada. É	fácil	gostar de cada actividade [→]
contexto	, eu achando que seria mais	fácil	do que na
contexto	tinha entendido que de Erasmus era	fácil	engatar com as meninas
contexto	em mim, que será mais	fácil	no campo.
contexto	diferente aqui e também bem mais	fácil	. Tenho muitas horas
contexto	Encontrar um emprego é provavelmente mais	fácil	na cidade mas
contexto	há poluição. vida é	fácil	mas no ponto
contexto	ano. A razão é muito	fácil	; morar no
contexto	cidade e lá é mais	fácil	escapar o estresse de
contexto	próprio meio de transporte, é	fácil	ficar isolado e pode
contexto	cidade esta é também acessível mais	fácil	e quando o tempo
contexto	muito saudável. E é mais	fácil	encontrar ambiente natural,
contexto	cadeiras □ O primeiro semestre parece mais	fácil	do que este
contexto	; para mim não foi nada	fácil	a primeira viagem sozinha
contexto	Avaliú muito, que há bastante	fácil	encontrar e visitar espectáculos
contexto	de Billy Joel porque não são	fácil	para aprender. Também
contexto	estudar? A vida não está	fácil	e arranjar colocação na
contexto	essa pergunta acaba por ser nada	fácil	. Começando com a
contexto	por isso que para mim é	fácil	acostumar-me à

[Descarregar resultados](#) - [Memorizar expressão de busca](#) - [Expressões CQL guardadas](#)

📌 Duas opções estão disponíveis:

i. Clicar em **ver** para lembrar os resultados obtidos anteriormente.

📌 Selecionar as pesquisas a comparar e clicar em **Comparar expressões de busca**.

📌 Depois, **Guardar** as expressões de pesquisa em formato **CSV**.

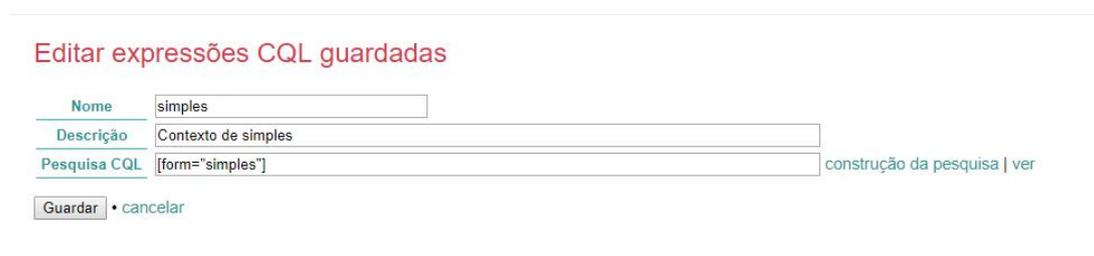
📌 **Nota:** Dependendo do *browser*, as expressões de pesquisa apenas ficam armazenadas temporariamente.



ii. Clicar em **editar** para preencher/alterar os campos **Nome** e **Descrição**.

📌 Preencha o nome com a palavra/expressão pesquisada.

📌 Finalmente, **Guardar** as expressões de pesquisa de acordo com os dados preenchidos.



📌 Pode, posteriormente, **Comparar expressões de busca** guardadas, selecionando-as.

Expressões CQL guardadas

Expressões CQL provisoriamente guardadas

- %5Bform%3D%22simples%22%5D+ [editar](#) [ver](#)
- fácil [editar](#) [ver](#) Contexto de fácil
- simples [editar](#) [ver](#) Contexto de simples

[Comparar expressões de busca](#)

De seguida, pode guardar esta comparação de pesquisas e voltar a pesquisar as expressões guardadas clicando em **Search for**.

Nota: Algumas funcionalidades ainda não estão disponíveis nesta área.

Comparação de pesquisas

Gráfico: Tabela | Guardar: [selecionar]

Search for Expressão de busca = fácil

Expressão de busca	Contagem
fácil	0
simples	0

[Ajuda](#)

**ANEXO II - FAQ E RESPOSTAS (PDF) DISPONÍVEIS NA PÁGINA
DE PESQUISA DO CORAL-CO**

1. Quais as anotações disponíveis para cada produção?

📖 A plataforma disponibiliza, para cada produção, as seguintes anotações:

Opções de representação

Powered by TEITOK
© Maarten Janssen,
2014-

Texto: **Transcrição** Forma do aluno - Mostrar: Cores - Etiquetas: Classe morfosintática Lema Áudio

i. Transcrição

📖 A *Transcrição* corresponde à produção original do aprendente, respeitando as suas hesitações, correções ou reformulações.

```
INF - Eh na primeira imagem há um //  
INF - É [uma] / uma história que / acho dá conta //  
INF - Muito países //  
INF - Eh // hhh //  
INF - Na primeira imagem hh há [um] um ladrão / Eh que Eh //  
INF - Depois na segunda vã fazer uma / rapina acho Uhm não sei como se diz a uma senhora / Eh //  
INF - Na terceira vã a roubar [a] a senhora vã deixar [a] a bolsa porque [...] quer ficar Ah tranquila Eh //  
INF - Na quarta imagem a polícia / Ah / fica atrás [de] <[...]> de um ladrão Eh //  
INF - E assim Ah a senhora fica [...] //  
INF - Pode tirar uma outra vez [a] / a sua / malas / Ah coisa [que] / que fá [na] na quinta imagem / onde //  
INF - O ladrão se volta Eh / e vê <o> a policia // na sexta imagem o ladrão Ah e a policia //  
INF - Se // s@ se podem ver são dois amigos / Eh / e na terceira v@ vã //  
INF - Abbraccio não sei como se diz / Ah dá um abraço e a senhora / Ah está um pouco surpreendida //  
INF - Já [na] / na outra imagem a senhora / vai embora porque depois [na] <nove> na nona imagem Ah / vã fazer a denúncia <de> [da] da situação //  
INF - Depois está o processo // o ladrão e a policia es@ [o] / o homem da policia estão / preocupados //  
INF - Ma@ Ah <[...]> [vã] tudo vã resolver-se porque também o / juiz é amigo do ladrão e <da> / do homem da policia sim //  
INF - Está tudo risolto como sempre //  
INT - Muito obrigada //
```

📖 Na fase de transcrição das produções orais, adotou-se o seguinte código de transcrição:

Convenções adotadas na transcrição	
/	Pausa curta
//	Pausa longa
Uhm Ah Eh Oh	Hesitações / Pausas preenchidas
hhh	Risos
xxx@	Truncamento
[xxx]	Repetição
<xxx>	Reformulação
[...]	Palavra ou segmento ininteligível
xxx	Estrangeirismo
xxx	Pouco claro / Existência de dúvida
Eva João	Nome genérico para mulheres Nome genérico para homens (Ocultação de elementos passíveis de reconstituir a identidade do falante)
INT (Tarefa 3) :XXX (Tarefa 2)	Segmentos onde se pode ouvir a voz do entrevistador
Bater de dedos	Cinesia

(Em alternativa, consultar [aqui](#))

ii. Forma do aluno

📄 É a versão “final” do aprendente, *limpa* de reformulações, repetições ou truncamentos.

📄 **Nota:** Apesar de ser visível o código a que se referem estas formas, elas não estão visíveis.

Opções de representação

Texto: - Mostrar: - Etiquetas:

▶ 0:00 / 1:57 🔊 ⋮

INF - *Eh* na primeira imagem há um //
 INF - É *[]* / uma história que / acho dá conta
 INF - Muito países
 INF - *Eh // nhã //*
 INF - Na primeira imagem *hã* há *[]* um ladrão / *E* que *Eh //*
 INF - Depois na segunda *vã* fazer uma / rapina acho *Uhm* não sei como se diz a uma senhora / *Eh //*
 INF - Na terceira *vã* a roubar *[]* a senhora *vã* deixar *[]* a bolsa porque [...] quer ficar *Ah tranquila Eh //*
 INF - Na quarta imagem a polícia / *Ah* fica atrás *[]* <> de um ladrão *Eh //*
 INF - E assim *Ah* a senhora fica [...] //
 INF - Pode tirar uma outra vez *[]* / a sua / malas / *Ah* coisa *[]* / que fá *[]* na quinta imagem / onde //
 INF - O ladrão se volta *Eh* / e vê <> a polícia // na sexta imagem o ladrão *Ah* e a polícia //
 INF - Se // @ se podem ver são dois amigos / *Eh* / e na terceira @ *vã* //
 INF - *Abbraccio* não sei como se diz / *Ah* dá um abraço e a senhora / *Ah* está um pouco surpreendida //
 INF - Já *[]* / na outra imagem a senhora / vai embora porque depois *[]* <> na nona imagem *Ah* / *vã* fazer a denúncia <> *[]* da situação //
 INF - Depois está o processo // o ladrão e a polícia @ *[]* / o homem da polícia estão / preocupados //
 INF - @ *Ah* <> *[]* tudo *vã* resolver-se porque também o / juiz é amigo do ladrão e <> / do homem da polícia sim
 INF - Está tudo *risolto* como sempre
 INT - Muito obrigada //

iii. Cores

- ✎ Embora esta opção de representação do texto exista por pré-definição na plataforma, neste caso não se aplica, porque as cores apenas foram usadas em algumas convenções de transcrição e não como critério para visualização dos dados.

iv. Classe morfofossintática

- ✎ Cada palavra tem uma etiqueta que a identifica quanto à sua classe morfofossintática, subclasse e flexão nominal (número e género) e verbal (tempo e modo).
- ✎ O pontuação de interrogação, o único sinal de pontuação utilizado na transcrição, também está anotado com uma etiqueta de natureza morfofossintática.

F1	Boa	tarde	desculpe	queria	um	copo	de	água	por	favor	//																
	Adv	Nome	Verbo	Verbo	Det	Nome	Prep	Nome	Prep	Nome																	
I1	Quero	sumo	//																								
	Verbo	Nome																									
I2	Queres	ir	ao	cinema	comigo	hoje	?	por	favor	//																	
	Verbo	Verbo	Prep+Det	Nome	Prep	Adv	Pontuação	Prep	Nome																		
F2	Boa	tarde	desculpe	queria	perguntar	uma	coisa	pequena	desculpa	?	posso	pode	ir	à	minha	pequena	feira	com	os	meus	pais	a	minha	família	faça	favor	//
	Adv	Nome	Verbo	Verbo	Verbo	Det	Nome	Adj	Verbo	Pontuação	Verbo	Verbo	Verbo	Prep+Det	Det	Adj	Nome	Prep	Det	Det	Nome	Det	Nome	Verbo	Verbo	Nome	
I3	Oh	meu	Deus	tu	és	um	idiota	//																			
	Interjeição	Det	Nome	Nome	Verbo	Det	Nome																				
F3	Desculpe	queria	saber	como	posso	fazer	agora	porque	tenho	um	compromisso	agora	e	não	tenho	uma	carta	como	é	possível	?	o	que	//			
	Verbo	Verbo	Verbo	Prep	Verbo	Verbo	Adv	Conj	Verbo	Det	Nome	Verbo	Conj	Adv	Verbo	Det	Nome	Adv	Verbo	Adj	Pontuação	Det	Conj				
I4	Muito	obrigada	fofinha	//																							
	Adv	Adv	Adj																								
F4	Muito	obrigada	foi	muito	gentil	não	sei	muito	muito	muito	obrigada	//															
	Adv	Adv	Verbo	Adv	Adj	Adv	Verbo	Adv	Adv	Adv	Adv																
F5	Desculpa	peço	desculpa	porque	foi	terrível	desculpa	mas	linha	um	problema	então	lamento	muito	//												
	Verbo	Verbo	Verbo	Conj	Verbo	Adj	Verbo	Conj	Verbo	Det	Nome	Adv	Verbo	Adv													
I5	Boa	tarde	desculpa	tenha	um	problema	com	as	minhas	companheiras	de	casa	desculpa	//													
	Adv	Nome	Verbo	Verbo	Det	Nome	Prep	Det	Nome	Nome	Prep	Nome	Verbo														
F6	Fico	feliz	por	esse	resultado	muitos	parabéns	os	melhores	cumpri	cumprimentos	//															
	Verbo	Adj	Prep	Det	Nome	Det	Nome	Det	Adj	Verbo	Nome																
I6	Muitos	parabéns	tu	és	a	melhor	menina	do	mundo	melhores	beijinhos	//															
	Det	Nome	Prep	Verbo	Det	Adj	Nome	Prep+Det	Nome	Verbo	Nome																

v. Lema

- 📖 A cada *token*, isto é, a cada unidade de significado, corresponde um lema.
- 📖 Quanto à lematização, deve ter-se em conta o seguinte:
 - 📖 As palavras compostas correspondem a apenas um lema (cf. exemplo: *fim de semana*);
 - 📖 As palavras contraídas, como é o caso das preposições com os determinantes, e os verbos pronominais (verbo+clítico) correspondem a dois lemas (cf. exemplo: *comprei-o*);
 - 📖 O ponto de interrogação também foi lematizado;
 - 📖 Não foram lematizados os seguintes segmentos:
 - 📖 truncamentos, repetições, reformulações, segmentos ininteligíveis ou pouco claros. ou estrangeirismos não foram lematizados;
 - 📖 segmentos extralinguísticos: pausas (curta e longa), pausas preenchidas, risos, cinesia.

F1	Boa tarde	desculpe	queria	um	copo	de	água	por	favor	//																										
I1	Quero	sumo	//																																	
I2	Queres	ir	ao	cinema	comigo	hoje	?	por	favor	//																										
F2	Boa tarde	desculpe	queria	perguntar	uma	coisa	pequena	desculpa	[pode]	pode	ir	à	minha	pequena	feira	com	os	meus	pais	a	minha	família	faça	favor	//											
I3	Oh	meu	Deus	tu	és	um	idiota	//																												
F3	Desculpe	queria	saber	como	posso	fazer	agora	porque	tenho	um	compromisso	agora	[e]	não	tenho	[a]	carta	como	[é]	como	é	possível	?	<Que>	[o]	que	posso	fazer	agora	?	pra	mim	é	muito	importante	//
I4	Muito	obrigada	fofinha	//																																
F4	Muito	obrigada	foi	muito	gentil	não	sei	muito	muito	muito	obrigada	//																								
F5	Desculpa	peço	desculpa	porque	foi	terível	desculpa	mas	tinha	um	problema	então	lamento	muito	//																					
I5	Boa tarde	desculpa	tinha	um	problema	com	as	minhas	companheiras	de	casa	desculpa	//																							
F6	Fico	feliz	por	esse	resultado	muitos	parabéns	os	melhores	cumprimentos	//																									
I6	Muitos	parabéns	tu	és	a	melhor	menina	do	mundo	muito	m@	beijinhos	//																							

vi. Áudio

📖 Ao seleccionar a opção Áudio, é possível ver o símbolo  antes de cada segmento textual.

▶ 0:00 / 1:37 🔊 ⋮

F1 - ▶ Boa tarde / desculpe queria um copo de água por favor //

I1 - ▶ Quero sumo //

I2 - ▶ Queres ir ao cinema comigo hoje? / por favor //

F2 - ▶ Boa tarde / desculpe queria perguntar uma coisa pequena desculpa [pode] pode ir à minha pequena festa com os meus pais a minha família faça favor //

I3 - ▶ Oh meu Deus tu és um idiota //

F3 - ▶ Desculpe queria saber como posso fazer agora porque tenho um compromisso agora [e] e não tenho [a] carta [como é] como é possível? <Que> [o] o que posso fazer agora? pra mim é muito importante //

I4 - ▶ Muito obrigada / fofinha //

F4 - ▶ Muito obrigada / foi muito gentil não sei muito muito muito obrigada //

F5 - ▶ Desculpa peço desculpa porque foi terrível desculpa mas tinha um problema então lamento muito //

I5 - ▶ Boa tarde / desculpa tinha um problema com as minhas companheiras de casa desculpa //

F6 - ▶ Fico feliz por esse resultado / muitos parabéns os melhores cumprimentos //

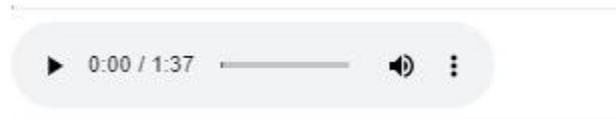
I6 - ▶ Muitos parabéns / tu és a melhor menina do mundo muito m@ beijinhos //

[Descarregar texto](#) • [Representação da onda sonora](#)

📖 Em relação à segmentação das produções é preciso saber que cada segmento foi delimitado em função das seguintes opções:

- 📖 para a tarefa 2, de acordo com a produção dos diferentes atos ilocutórios;
- 📖 para a tarefa 5, como consiste na leitura de um texto escrito, a segmentação foi feita respeitando a pausa assinalada, ortograficamente, com um ponto final;

- ✍ para as tarefas 6 e 7, por palavra;
- ✍ para as tarefas 1 (entrevista) e 3 (conto de uma história a partir de suporte pictórico), com base nas pausas longas que estruturam o discurso.
- ✍ Se pretender ouvir a produção integral, sem interrupções clique no símbolo ► na figura que precede a transcrição.



- ✍ No final da transcrição, é possível seleccionar outra opção de visualização da produção, clicando em **Representação da onda sonora**.

Representação da onda sonora
001_B1_T2

Velocidade: 100% 2.898 / 1.37.152 Zoom: 100 pps

F1 - Boa tarde / desculpe queria um copo de água por favor //

I1 - Quero sumo //

I2 - Queres ir *Ah* ao cinema comigo hoje? / por favor //

F2 - Boa tarde / *Ah Ah* desculpe queria perguntar uma coisa pequena desculpa / [pode] pode ir à minha pequena festa com <ou> os meus pais a minha família faça favor *nhh* //

I3 - Oh meu Deus tu és um idiota //

F3 - Desculpe queria saber *Ah Ah* como posso fazer agora porque tenho / *nhh* um compromisso agora [e] e não tenho [a]; *Uhm* / a carta [como é] como é possível? <Que> / o que posso fazer agora? / pra mim é muito importante //

I4 - Muito obrigada / fofinha *Ah nhh* //

F4 - Muito obrigada :*Uhm* / foi muito gentil *nhh* / *Ah Ah* não sei *m@* muito muito obrigada //

F5 - Desculpa peço desculpa porque / foi terrível desculpa / ;*Uhm* mas *nh* tinha um problema <um> *En* então *Ah* / <sou> *Ah* / lamento muito *nhh* //

I5 - Boa tarde desculpa tinha um problema com as minhas companheiras de casa *nhh* desculpa //

F6 - Fico feliz / por esse resultado / muitos parabéns *Et* os melhores *cumpri* cumprimentos //

I6 - Muitos parabéns *Ah* / tu és a melhor menina do mundo / <muito> *m@* beijinhos //

- ✍ Esta opção permite:

i. ouvir os segmentos textuais, um a um, clicando sobre cada um deles;

ii. ouvir a produção integral, clicando em play ►

iii. parar a reprodução, puxar para trás ou para a frente, as vezes necessárias, utilizando os botões de comando: ◀▶▶

iv. escolher um segmento clicando diretamente na onda.

2. Como identificar os códigos dos estímulos a partir dos quais as produções orais foram obtidas?

- 📄 As tarefas de produção oral foram obtidas a partir dos estímulos descritos na *Metodologia* do *corpus* ([aqui](#)).

Inquérito para recolha dos dados

Os dados orais foram recolhidos pela aplicação de um Inquérito com a seguinte estrutura:

Secção	Natureza dos dados	Tarefa
1	Produção oral	1 - Entrevista semiestruturada
2		2 - Elicitação de atos ilocutórios
		3 - Construção de um texto narrativo a partir de uma sequência de imagens
		4 - Nomeação de figuras a partir de um suporte pictórico
3	Leitura oral	5 - Leitura de texto
		6 - Leitura de listas de palavras
		7 - Leitura de listas de palavras

3. Como podem ser visualizados os resultados de pesquisa? *Há opções.*

Os resultados da pesquisa podem ser visualizados de duas formas diferentes.

📌 Preencher os campos da construção da pesquisa, e, de seguida, clicar em **opções**.

i. Os campos das **Opções de busca** estão pré-preenchidos por defeito porque os resultados são, por definição, apresentados em linha de contexto (*Key Word in Context*), correspondendo à combinação mais curta de 5 palavras.

📌 Quando uma pesquisa é criada, os resultados são sempre apresentados desta forma, automaticamente, não sendo necessário preencher qualquer campo.

Pesquisa no corpus

CQP Query: [lemma = "ladrao"] within text [Pesquisar] construção da pesquisa | ver | opções

Opções de busca

Tipo de representação visual: KWIC Context

Tamanho do contexto: [5] palavras

Ordenar por: [Palavra]

Estratégia de combinação: [Combinação mais longa]

Construção de pesquisa

Pesquisa do texto

Forma do aluno: [igual a]

Forma normalizada: [igual a]

Classe morfosintática: [construção de etiquetas]

Lema: [igual a] ladrao

[Adicionar token]

Pesquisa do documento

Informante: [selecionar]

Tarefa: [selecionar]

Proficiência: [selecionar]

Nacionalidade: [selecionar]

Língua materna: [selecionar]

Outra pesquisa

Tarefa 2 (subtarefas): [selecionar]

Formal / Informal: [selecionar]

Ato ilocutório: [selecionar]

Search within: [Text]

[Pesquisar] cancelar | ajuda

Pesquisa no corpus

COP Query: [lema = "ladrao"] within text [Pesquisar] construção da pesquisa | ver | opções

Visualização da pesquisa COP

1 Lema = ladrao

161 resultados • A mostrar 0 - 100 (seguintes)

Texto: [Transcrição] [Forma do aluno]

Etiquetas: [Classe morfosintática] [Lema]

contexto E um crime Um ladrão Adrao ? Ladrão [[Adrao] [.] adrao An # Salta
contexto E de repente aparece um Ladrão / com uma arma que diz
contexto gente incluindo a policia e # Ladrão / [.] toda gente foram [[.] Todos amigos
contexto crime # Um ladrão / Adrao? # Ladrão # [Adrao] [.] adrao An # Salta? salta
contexto palavra Um ladrão Ladrão s @ An encontrou um ladrão # An # Logo
contexto ladrão # Adrao? # Ladrão # [Adrao] [.] adrao / An # Salta? salta para a
contexto primeira imagem há um ladrão / E que # Depois na segunda
contexto atrás [le] - # de um ladrão E # E assim a senhora fica
contexto quinta imagem onde # O ladrão se volta / e vê >
contexto sexta imagem o ladrão An e a policia / Se # s @
contexto Depois está o processo # o ladrão e a policia s @ [o]
contexto juiz é amigo do ladrão e da / do homem
contexto pode prender # o ladrão / mas depois Lhm / [o] s @
contexto depois # [o] [o] # o ladrão e o agente # Lhm # Lhm foram
contexto de agente e do ladrão [A] his @ A historia não
contexto foi assaltada pelo um ladrão # Que tentou roubar [a sua]
contexto policia conhecia > ao ladrão então eles # # Eles cumprimentaram e
contexto policia não fez nada contra ladrão ni tentou é deter An # aprisionario
contexto ser também conhecido do ladrão e da policia
contexto foi roubar # # pelo um ladrão # E roubar [a] a sua
contexto a a policia e o ladrão são amigos # # Portanto a <mulh >
contexto o que aconteceu E / depois / Lhm # O ladrão [e] -a- e o policia
contexto coisas de valor para o ladrão # E Lhm / nesta altura | De

ii. Opcionalmente, os resultados podem ser apresentados por contexto, numa combinação mais longa de (e até 100) *tokens*.

☞ Para isso, selecionar os campos **Context** e **Mostrar contexto** e escolher o número de *tokens* que pretende.

☞ Por fim, clicar em **Pesquisar**.

Pesquisa no corpus

COP Query: [lema = "ladrao"] within text [Pesquisar] construção da pesquisa | ver | opções

Opções de busca

Tipo de representação visual: KWIC Context

Mostrar contexto: Tokens: 30 Utterance Search

Ordenar por: [Palavra]

Estratégia de combinação: [Combinação mais longa]

Construção de pesquisa

Pesquisa do texto

Forma do aluno [igual a]

Forma normalizada [igual a]

Classe morfosintática [construção de etiquetas]

Lema [igual a] ladrao

Adicionar token

Pesquisar cancelar ajuda

Pesquisa do documento

Informante [selecionar]

Tarefa [selecionar]

Proficiência [selecionar]

Nacionalidade [selecionar]

Língua materna [selecionar]

Outra pesquisa

Tarefa 2 (subtarefas) [selecionar]

Formal / Informal [selecionar]

Ato ilocutório [selecionar]

Search within: [Text]

Pesquisa no corpus

CQP Query: [lemma = "ladrao"] within text construção da pesquisa | ver | opções

Visualização da pesquisa CQP

Lema = ladrao

161 resultados • A mostrar 0 - 100 (seguintes)

Texto:

Etiquetas:

contexto Havia uma mulher / Sim / que anda na rua / Um ladrão / Adrao? / Ladrão / Adrao / adrao / Salta? / salta para a frente dela / como / Como se diz?

contexto Uma professora / ou / sim não / Um juiz / está a anda / E / de repente / aparece um / Ladrão / com uma arma que diz / Que diz / à juiza de / lhe dar / a mala / A sua mala / e / com medo ela /

contexto vamos fazer um corte / Durante ele espera / ela espera / para vítima / como vítima para / Fazer / Um apresentação explicação de que aconteceu / quando juris chegou / toda gente / incluindo a polícia / e / Ladrão / toda gente foram / Todos amigos / a problema / A senhora pensava / é / mais grande que pensava

contexto Havia uma mulher / Sim / que anda na rua / Um ladrão / Adrao? / Ladrão / Adrao / adrao / Salta? / salta para a frente dela / como / Como se diz?

contexto Uma pistola / Pistola / uma pistola / ela pidi para / darfe / sua mala / com nhedo

contexto um dia / a senhora Ana / está na rua e encontrou um / Tirou um / Esqueci-me da palavra / Um ladrão / Ladrão / encontrou / um ladrão / Logo que a senhora Ana / viu o ladrão / ele / Tirou um / Puntou a revolver / para a senhora / e / disse / Engreque-me

contexto Havia uma mulher / Sim / que anda na rua / Um ladrão / Adrao? / Ladrão / Adrao / adrao / Salta? / salta para a frente dela / como / Como se diz?

contexto Uma pistola / Pistola / uma pistola / ela pidi para / darfe / sua mala / com nhedo

contexto primeira imagem há um / É / uma história que / acho dá conta / Muito países / Na primeira imagem / há um / ladrão / que / Depois na segunda / vai fazer uma / rapina acho não sei como se diz e uma senhora / Na terceira / a senhora / vai deixar / a /

contexto terceira / a roubar / a senhora / vai deixar / a bolsa porque / quer ficar / tranquila / Na quarta imagem a polícia / fica atrás / de um / ladrão / E assim / a senhora fica / Pode tirar uma outra vez / a / suas / malas / coisa / que / na / quinta imagem / onde / O ladrão se volta / e / vê / de / de um / ladrão / E assim / a senhora fica / Pode tirar uma outra vez / a / suas / malas / coisa / que / na / quinta imagem / onde / O ladrão se volta / e / vê / a / polícia / Na sexta imagem o ladrão / e a / polícia / Se / se / se podem ver são dois amigos / e / na terceira /

contexto outra vez / a / suas / malas / coisa / que / na / quinta imagem / onde / O ladrão se volta / e / vê / a / polícia / Na sexta imagem o ladrão / e a / polícia / Se / se / se podem ver são dois amigos / e / na terceira /

contexto outra imagem a senhora / vai embora porque depois / na nona imagem / vai fazer a denúncia / da / da situação / Depois está o processo / o / ladrão / e a / polícia / o / homem da polícia estão / preocupados / tudo / vai resolver-se porque também o / juiz / é amigo do ladrão e /

iii. Uma terceira opção para visualização dos dados está disponível. Os resultados podem ser apresentados nos segmentos textuais em que estão inseridos.

☞ Para isso, selecionar nas **Opções de busca**, a opção **Utterance search** e **Mostrar contexto**.

☞ Por fim, clicar em **Pesquisar**.

Pesquisa no corpus

CQP Query: construção da pesquisa | ver | opções

Opções de busca

Tipo de representação visual: KWIC Context

Mostrar contexto: Tokens: 30 Utterance Search

Ordenar por: Palavra

Estratégia de combinação: Combinação mais longa

Construção de pesquisa

Pesquisa do texto

Forma do aluno: igual a

Forma normalizada: igual a

Classe morfosintática: construção de etiquetas

Lema: igual a ladrao

Pesquisa do documento

Informante: [selecionar]

Tarefa: [selecionar]

Proficiência: [selecionar]

Nacionalidade: [selecionar]

Língua materna: [selecionar]

Outra pesquisa

Tarefa 2 (subtarefas): [selecionar]

Formal / Informal: [selecionar]

Ato ilocutório: [selecionar]

Search within: Text

Pesquisa no corpus

CQP Query: [lemma = "ladrão"] within text [Pesquisar] construção da pesquisa | ver | opções

Visualização da pesquisa CQL

Lema = ladrão

161 resultados • A mostrar 0 - 100 (seguintes)

Texto: [Transcrição] [Forma do aluno]

Etiquetas: [Classe morfosintática] [Lema]

contexto Adrão? // |

contexto Ladrão com uma arma que diz [que] [que] que / Uhm // |

contexto Ladrão [...] toda gente foram // |

contexto Ladrão // |

contexto Ladrão s@ Ah / encontrou um ladrão // Ah // |

contexto [Adrão] [...] adrão / Ah // |

contexto Na primeira imagem nra há [um] um ladrão / En que Ah // |

contexto Na quarta imagem a polícia / Ah / fica atrás [de] - de um ladrão Ah // |

contexto O ladrão se volta e vê a polícia / na sexta imagem o ladrão e a polícia // |

contexto O ladrão se volta e vê a polícia / na sexta imagem o ladrão e a polícia // |

contexto Depois está o processo / o ladrão e a polícia / o homem da polícia estão preocupados // |

contexto [Mig] - tudo vai resolver-se porque também o juiz é amigo do ladrão e do homem da polícia sim |

contexto Uhm feliz que ele apareceu porque / o agente pode / prender / o / ladrão |

contexto Uhm mas depois Uhm / o / o ladrão e o agente / Uhm / Uhm // |

contexto Uhm / de agente e do ladrão |

contexto É uma senhora que estava a caminhar pela rua a passear / e foi assaltada pelo um ladrão // |

contexto Mas / foi engraçado porque o polícia / conhecia / ao ladrão então eles // |

contexto Por causa de isso porque a polícia / não fiz nada contra ladrão ni tentou / é deter / ah / aprisionarlo // |

contexto 30.04.2014 Portugal Coimbra tokenized using xmltokenize.ptTagged with Freeling | É uma senhora que estava a caminhar pela rua a passear / e foi assaltada pelo um ladrão // | Que tentou roubar [a sua] / [a sua mala] [a] s@ s@ [sua] / sim a sua mala | E / Ah / depois / Ah / apresentou-se -a- / o polícia // | Mas // / foi engraçado porque o polícia / conhecia / ao ladrão / então eles // | Eles cumprimentaram e ficaram felizes por encontrar-se / e a senhora ficou chateada // | Por causa de isso porque a polícia / não fiz nada contra ladrão ni tentou / é deter / Ah / aprisionarlo // | E então depois / Uhm / a senhora / foi para Ah [a] a comissaria // | A delegacia de polícia / e ni / [a] a demandar [os] [os] [os] / os factos que aconteceram / há um bocadinho // | E depois [da] / [da] comissaria / da delegacia / Foram para o juízo / Uhm / e no momento de começar o juízo quando entrou o juiz |

contexto Uma mulher foi roubar / pelo um ladrão |

📖 Após a exibição dos resultados, passar com o curso por cima das palavras para uma rápida observação da etiquetagem das palavras.

Pesquisa no corpus

CQP Query: [lemma = "ladrão"] within text [Pesquisar] construção da pesquisa | ver | opções

Visualização da pesquisa CQL

Lema = ladrão

161 resultados • A mostrar 0 - 100 (seguintes)

Texto: [Transcrição] [Forma do aluno]

Etiquetas: [Classe morfosintática] [Lema]

contexto E um / crimene // | Um ladrão // | Adrão ? // | Ladrão // | [Adrão] / [...] adrão / Ah // | Salta

contexto E / de repente Ah aparece um // | Ladrão / com uma arma que diz

contexto gente / incluindo la polícia / e // | Ladrão / [...] toda gente foram // | [...] Todos amigos

contexto crimene // | Um ladrão // | Adrão? // | Ladrão // | [Adrão] / [...] adrão / Ah // | Salta? / salta

contexto palavra / Um ladrão / Ladrão s@ Ah / encontrou um ladrão // Ah // | Logo

contexto ladrão // | Adrão? // | Ladrão // | [Adrão] / [...] adrão / Ah // | Salta? / salta para a

contexto primeira imagem nra há [um] um ladrão / En que Ah // | Depois na segunda

contexto atrás [de] - de um ladrão En // | E assim Ah / a senhora fica

contexto quinta imagem / onde // | O ladrão se volta / e vê ->

contexto sexta imagem o ladrão Ah / e a polícia // | Se // s @

contexto Depois está o prc imagem

contexto juiz é

contexto pode / p@ prent

contexto depois Uhm / [o]

contexto de aç

contexto foi assaltada peio um ladrão // | Que tentou roubar [a sua] |

contexto polícia / conhecia <o> ao ladrão / então eles Ah // | Eles cumprimentaram e

contexto polícia / não fiz nada contra ladrão ni tentou / é deter Ah / aprisionarlo

contexto ser também conhecido do ladrão e da polícia

contexto foi roubar Ah / pelo um ladrão // | E / roubar [a] a sua

contexto a a polícia e o ladrão são amigos nra / Ah Uhm // | Portanto a <mulhr >

contexto o que aconteceu / En / depois / Uhm // | O ladrão [e] <a> e o polícia

contexto coisas de valor para / o ladrão // | E Uhm / nesta altura / De

📖 Para consultar a anotação linguística detalhada, aceder à produção do aprendente na íntegra, clicando em **contexto**, no início de cada linha de ocorrência.

4. Adicionar *token*: como utilizar este critério de pesquisa? Se faz favor!

- ✍ A cada *token* corresponde um lema.
- ✍ Para pesquisar uma determinada expressão, constituída por mais do que uma palavra, escrever no campo **Lema** a primeira palavra da sequência e clicar em **Adicionar token**.
- ✍ Repetir o processo adicionando outros lemas/*tokens*, sempre por ordem, até obter a expressão pretendida.
- ✍ Sempre que adiciona um *token*, pode observar-se o conjunto de lemas adicionados pela ordem em que serão pesquisados.

Pesquisa no corpus

CQP Query: construção da pesquisa | ver | opções

Construção de pesquisa

1	Lema = se	2	Lema = fazer	3	Lema = favor
---	-----------	---	--------------	---	--------------

Pesquisa do texto

Forma do aluno: igual a

Forma normalizada: igual a

Classe morfosintática: construção de etiquetas

Lema: igual a

Pesquisa do documento

Informante: [selecionar] ▼

Tarefa: [selecionar] ▼

Proficiência: [selecionar] ▼

Nacionalidade: [selecionar] ▼

Língua materna: [selecionar] ▼

Outra pesquisa

Tarefa 2 (subtarefas): [selecionar] ▼

Formal / Informal: [selecionar] ▼

Ato ilocutório: [selecionar] ▼

Search within: Text ▼

Pesquisa no corpus

CQP Query: [lema = "se"] [lema = "fazer"] [lema = "favor"] within text construção da pesquisa | ver | opções

Visualização da pesquisa CQL

1	Lema = se	2	Lema = fazer	3	Lema = favor
---	-----------	---	--------------	---	--------------

23 resultados

Texto:

Etiquetas:

contexto		Se faz favor	queria um copo de água
contexto	Era uma água	se faz favor	Podés ir buscar meu sumo
contexto	☺ ☺ o sumo?	se faz favor	então cara professora a senhora
contexto	Uma garrafa de água	se faz favor	Sm Chega-me um copo de
contexto	um copo de sumo	se faz favor	Senhora professora não se importa
contexto	olhe	se faz favor	eu quella um copo de
contexto	aquele copo de sumo	se faz favor	Professor deve ser muito formal
contexto	fazer esta [...] recomendação pra mim	se faz favor	eu realmente preciso dele
contexto	importava de ☺ ☺ fazer agora?	se faz favor	O pá fico muito zangada
contexto	favor Podés passar aquele sumo	se faz favor	Gostaria de convidar se
contexto	passar o sumo para mim	se faz favor	Olá professora muito obrigado
contexto	posso receber o copo?	se faz favor	☺ ☺ pode passar o sumo de
contexto	uma garrafa de água fresca	se faz favor	☺ ☺ / passa-me por favor @ o
contexto	queria uma garrafa de água	se faz favor	com gelo Olha pode dar-me
contexto	pode dar-me o sumo	se faz favor	? A senhora eu vou
contexto	empresa Ah ☺ ☺ já estou atrasado mas	se faz favor	pode me escrever ☺ ☺ @ o cartão
contexto	uma água mineral com gás	se faz favor	Por favor se faz favor
contexto	se faz favor Por favor	se faz favor	me passa o sumo Diria
contexto	Posso ter uma água	se faz favor	? Passa-me o sumo
contexto	? Passa-me o sumo	se faz favor	☺ ☺ A gente é muito ☺ ☺ informal
contexto	água Dá-me o sumo	se faz favor	Professor podia ir ☺ ☺ pala qualquer

- 📌 **Nota:** Quando combinamos o preenchimento do campo **Lema** com o campo **Classe morfossintática**, estamos a pesquisar uma palavra específica que pertence a uma determinada classe.

Pesquisa no corpus

CQP Query: [construção da pesquisa](#) | [ver](#) | [opções](#)

Construção de etiquetas: Classe morfossintática

POS principal:

- Adjetivo
- Advérbio
- Conjunção
- Data
- Determinante
- Interjeição
- Nome
- Número
- Pontuação
- Preposição
- Pronome
- Verbo

Inserir | Acres

Construção

Pesquisa de

- Forma
- Forma no
- Classe morfossintática

Construção de etiquetas

Lema igual a

[cancelar](#) | [ajuda](#)

Pesquisa do documento

Informante

Tarefa

Proficiência

Nacionalidade

Língua materna

Outra pesquisa

Tarefa 2 (subtarefas)

Formal / Informal

Ato ilocutório

Search within:

Pesquisa no corpus

CQP Query: [construção da pesquisa](#) | [ver](#) | [opções](#)

Visualização da pesquisa CQL

1
Classe morfossintática = [starts with] N
and Lemma = roubo

8 resultados

Texto:

Etiquetas:

```

contexto      a este problema // | A este roubo // sim
contexto      rua / e // ela encontrou // o roubo // | E / o roubo // | O roubo
contexto      encontrou // o roubo // | E / o roubo // | O roubo tra@ trazia |a| a // é
contexto      roubo // | E / o roubo // | O roubo tra@ trazia |a| a // é mala?
contexto      e // | Mas / o roubo <não> / <nunca se> não q@ sabia que a policia
contexto      era policia e // | Quando o roubo // | [...] // | Turned his back? Turned
contexto      era / o // | <Amigo> Amigos com o roubo e // | A mulher está muito
contexto      juiz // | O juiz // <chegava> // <chegava> // chegou // | O roubo / era // | Uhm // | O juiz era / outro
  
```

[Descarregar resultados](#) - [Memorizar expressão de busca](#)

- 📌 Já quando queremos pesquisar uma palavra seguida de outra que pertence a determinada classe morfossintática, então, é necessário, depois de preencher o campo **Lema**, **Adicionar token** e só depois escolher a **Classe morfossintática**.

Pesquisa no corpus

COP Query: [Pesquisar] construção da pesquisa | ver | opções

Construção de etiquetas: Classe morfosintática

POS principal: [selecionar]

Inserir [Ativar] [Adjetivo] [Advérbio] [Conjunção] [Data] [Determinante] [Interjeição]

1 Lema = roubar [Nome]

Pesquisa de: [Portuação] [Forma] [Pronome] [Forma no Verbo]

Classe morfosintática: construção de etiquetas

Lema: igual a

Adicionar token

Pesquisar cancelar ajuda

Pesquisa do documento

Informante: [selecionar]

Tarefa: [selecionar]

Proficiência: [selecionar]

Nacionalidade: [selecionar]

Língua materna: [selecionar]

Outra pesquisa

Tarefa 2 (subtarefas): [selecionar]

Formal / Informal: [selecionar]

Ato ilocutório: [selecionar]

Search within: Text

Pesquisa no corpus

COP Query: [lema = "roubar" [pos = "N*"] within text] [Pesquisar] construção da pesquisa | ver | opções

Visualização da pesquisa CQL

1 Lema = roubar 2 Classe morfosintática = [starts with] N

1 resultados

Texto: [Transcrição] [Forma do aluno]

Etiquetas: [Classe morfosintática] [Lema]

contexto um homem # | Tenta @ saltar para / roubar dinheiro | E / uma mulher &# | Entregou a

5. Qual a funcionalidade de *A acrescentar ao lado*, na *Construção de etiquetas*? *Tenho tempo ...*

A opção **A acrescentar ao lado** permite fazer uma pesquisa inserindo uma classe morfosintática em alternativa à que foi inserida previamente, ou seja, permite pesquisar a ocorrência de diferentes classes em determinado contexto.

- ✎ Consideremos o exemplo do verbo *ter*: poder ser um verbo auxiliar ou um verbo pleno dependendo do contexto frásico em que ocorre.
- ✎ Esta opção vai permitir pesquisar, simultaneamente, dois dos contextos em que o verbo pode ocorrer, por exemplo, *ter+verbo // ter+nome*.
- ✎ Primeiro, preencha o campo **Lema** com a forma do infinitivo do verbo *ter* e clique em **Adicionar token**.

The screenshot shows the 'Pesquisa no corpus' interface. At the top, there is a 'CQP Query:' field and a 'Pesquisar' button. Below this is a 'Construção de pesquisa' dialog box. The dialog has a 'Lema = ter' field. Under 'Pesquisa do texto', there are fields for 'Forma do aluno' (igual a), 'Forma normalizada' (igual a), 'Classe morfosintática' (construção de etiquetas), and 'Lema' (igual a). There is an 'Adicionar token' button. On the right, 'Pesquisa do documento' includes dropdowns for 'Informante', 'Tarefa', 'Proficiência', 'Nacionalidade', and 'Lingua materna'. Below that, 'Outra pesquisa' includes dropdowns for 'Tarefa 2 (subtarefas)', 'Formal / Informal', and 'Ato ilocutório'. At the bottom of the dialog, there is a 'Search within:' dropdown set to 'Text' and buttons for 'Pesquisar', 'cancelar', and 'ajuda'.

- ✎ De seguida, clique em **construção de etiquetas > POS principal**, selecione a categoria **Nome** e clique em **Inserir**.

Pesquisa no corpus

CQP Query: construção da pesquisa | ver | opções

Construção de etiquetas: Classe morfosintática

POS principal: Nome

Tipo [qualquer]

Gênero [qualquer]

Número [qualquer]

Grau [qualquer]

Construção de pesquisa

Lema = ler

Pesquisa do texto

Forma do aluno igual a

Forma normalizada igual a

Classe morfosintática construção de etiquetas

Lema igual a

Pesquisa do documento

Informante [selecionar]

Tarefa [selecionar]

Proficiência [selecionar]

Nacionalidade [selecionar]

Língua materna [selecionar]

Outra pesquisa

Tarefa 2 (subtarefas) [selecionar]

Formal / Informal [selecionar]

Ato Ilocutório [selecionar]

Search within: Text

Pesquisa no corpus

CQP Query: construção da pesquisa | ver | opções

Construção de pesquisa

Lema = ler

Pesquisa do texto

Forma do aluno igual a

Forma normalizada igual a

Classe morfosintática construção de etiquetas N.*

Lema igual a

Pesquisa do documento

Informante [selecionar]

Tarefa [selecionar]

Proficiência [selecionar]

Nacionalidade [selecionar]

Língua materna [selecionar]

Outra pesquisa

Tarefa 2 (subtarefas) [selecionar]

Formal / Informal [selecionar]

Ato Ilocutório [selecionar]

Search within: Text

- 📄 Volte a **construção de etiquetas > POS principal**, selecione a categoria **Verbo** e clique em **Acrescentar ao lado**. A pesquisa está construída.
- 📄 Finalize clicando em **Pesquisar** para obter os resultados.

Pesquisa no corpus

CQP Query: construção da pesquisa | ver | opções

Construção de pesquisa

Lema = ler

Pesquisa do texto

Forma do aluno igual a

Forma normalizada igual a

Classe morfosintática construção de etiquetas N*IV*

Lema igual a

Pesquisa do documento

Informante [selecionar]

Tarefa [selecionar]

Proficiência [selecionar]

Nacionalidade [selecionar]

Língua materna [selecionar]

Outra pesquisa

Tarefa 2 (subtarefas) [selecionar]

Formal / Informal [selecionar]

Ato Ilocutório [selecionar]

Search within: Text

As estruturas *ter+nome* /*ter+verbo* surgem destacadas no contexto em que ocorrem.

Pesquisa no corpus

CQP Query: [lemma = "ter"] [pos = "N,V"] within text [Pesquisar] construção da pesquisa | ver | opções

Visualização da pesquisa CQL

Lema = ter | Classe morfosintática = N,V*

43 resultados

Texto | Transcrição | Forma do aluno

Etiquetas | Classe morfosintática | Lema

contexto	festa / [/ @ <=> que eu organizo? # [] <=>	Tem tempo	[] <=> [] que fomos ao cinema
contexto	aniversário? festa de aniversário # []	Tem tempo	para / vai contigo / ao
contexto	uma queixa e depois = tribunal # []	Temos <=> deve	ser o bandido # [] O polícia
contexto	una razão # [] # [] # [] Tem uma razão # []	Tenho razão	[] # [] # [] a professora # [] # [] não animada # [] Ela
contexto	dóis / # [] # [] Os dois vão ter # []	<=> teve em ter	# [] # [] # [] Mas o que
contexto	isso # [] e # [] # [] o ladrão / ficou # []	Teve medo	/ e -> / a senhora Susana # [] # [] Ganhou
contexto	um ladrão / E / o ladrão	tem arma	# [] E / # [] # [] qui tá / As coisas
contexto	praia. Como não	tem carro	, conhece bem o metro
contexto	hoje às seis? # [] # [] # [] # [] # [] # []	tem tempo	para vamos ao cinema
contexto	minha licenciatura >= sim então # [] se	tem tempo	está bem-vindo / para vir # [] Olá
contexto	final / de semana / você	tem tempo	/ para ir / à festa
contexto	e a professora agora ainda	tem tempo	/ para o trabalho? # [] Peço
contexto	yo qué fazemos / e / como /	temos = exemplos	de / trabalho / que = temos que
contexto	noite de amani /	temos festa	com amigos # [] = estamos # [] tipo agora
contexto	Agora é muito difícil porque	temos salas	# [] muito grande e / não podemos
contexto	e depois <=> embora os polícia	tenha chegado	# [] # [] aquele # [] polícia / Conhecia / este ladrão
contexto	ocupada hoje? porque eu	tenho @ bilhetes	para / ir ao [] / queres
contexto	obrigada pela flor # [] Não	tenho cara	para tanta vergonha não queria
contexto	hora <=> tentei / ligar mas = não	tenho contato	como a senhora / e desculpe
contexto	fui jantar e eu não	tenho dinheiro	= a minha telefone eu / não
contexto	noite? / Obrigada eu	tenho estudar	muito para / = finalmente exame # [] # [] que
contexto	todo trânsito <=> # [] Não sei / Não	tenho ideia	/ # [] não sei # [] Parabéns <=> muitos parabéns
contexto	fim <=> [] do curso eu	tenho jantar	com minha amigas e eu

Nota: Este tipo de pesquisa também pode ser útil no estudo de casos relacionados com (in)transitividade dos verbos, regência preposicional, colocações na frase.

6. Como pesquisar palavras contraídas? *Deste e daquele.*

📌 As palavras contraídas correspondem a dois lemas pertencentes a classes morfossintáticas distintas.

📌 Existem duas formas de pesquisar palavras contraídas:

i. Por **Lema**, obtendo todas as palavras que resultam da contração com o lema pesquisado;

ii. Por combinação de dois lemas, no caso de pretender pesquisar uma preposição específica contraída com uma palavra pertencente a uma determinada classe /subclasse.

Pesquisa no corpus

COP Query: [construção da pesquisa](#) | [ver](#) | [opções](#)

Construção de pesquisa

1	2
Lema = de	Lema = este

Pesquisa do texto

Forma do aluno	igual a	<input type="text"/>
Forma normalizada	igual a	<input type="text"/>
Classe morfossintática	construção de etiquetas	<input type="text"/>
Lema	igual a	<input type="text"/>

[cancelar](#) | [ajuda](#)

Pesquisa do documento

Informante	[selecionar]
Tarefa	[selecionar]
Proficiência	[selecionar]
Nacionalidade	[selecionar]
Língua materna	[selecionar]

Outra pesquisa

Tarefa 2 (subtarefas)	[selecionar]
Formal / Informal	[selecionar]
Ato ilocutório	[selecionar]

Search within:

7. Como pesquisar palavras hifenizadas? *Amigos de infância abraçam-se.*

Como o hífen não foi lematizado, são duas as opções de pesquisa possíveis para os dois tipos de palavras hifenizadas:

i. palavras compostas: são consideradas como apenas um lema e, por isso, devem ser pesquisadas por **Lema**.

Nota: Aguardamos a transcrição de todas as produções para observar a existência de palavras compostas no *corpus*.

ii. conjugação pronominal ou reflexa: deve considerar-se dois lemas (o verbo e o clítico).

Pesquisa no corpus

CQP Query: Pesquisar construção da pesquisa | ver | opções

Construção de pesquisa

1 Lema = abraçar 2 Lema = se

Pesquisa do texto

Forma do aluno igual a

Forma normalizada igual a

Classe morfosintática construção de etiquetas

Lema igual a

Adicionar token

Pesquisa do documento

Informante [selecionar]

Tarefa [selecionar]

Proficiência [selecionar]

Nacionalidade [selecionar]

Língua materna [selecionar]

Outra pesquisa

Tarefa 2 (subtarefas) [selecionar]

Formal / Informal [selecionar]

Ato /ocutório [selecionar]

Search within: Text

Pesquisar cancelar ajuda

Pesquisa no corpus

CQP Query: [lemma = "abraçar"] [lemma = "se"] within text Pesquisar construção da pesquisa | ver | opções

Visualização da pesquisa CQL

1 Lema = abraçar 2 Lema = se

8 resultados

Texto: Transcrição Forma do aluno

Etiquetas: Classe morfosintática Lema

contexto ele / An / A / aquele polícia aquele ladrão / abraçaram-se e também / cumprimentaram-se e

contexto temos / eles / An / / / não sei / eles / abraçaram-se / E / aqui / Não compreendo / aqui

contexto o seu amigo melhor / / eles abraçam-se e senhora estava muito irritada

contexto amigos / Amigos da infância / abraçam-se / E / a mulher fica com

contexto conhecem de toda a vida / abraçam-se / / A senhora / acha / isto insólito

contexto polícia / Eran dois velhos amigos / abraçam-se e a senhora / / vai-se

contexto para encontrar na rua abraçando-se não a eles / importa / / / o

contexto anos / / ficaram / Ambos ficaram contentes / / / Abraçaram-se / E / / isso realmente / / surpreendeu a

Nota: A pesquisa por **Pontuação > hífen** devolve, por defeito, ocorrências com travessão e não com hífen.

8. Como pesquisar sufixos e prefixos? *É possível!*

- Existe uma forma de pesquisar sufixos e prefixos que, não sendo exclusiva para este fim, é bastante produtiva.
- Ao preencher o campo **Lema**, selecionar primeiro **começa por / termina com** e depois escrever o prefixo / sufixo pretendidos.
- Da mesma forma, pode pesquisar um radical, selecionando a opção **contém** no **Lema**.
- O preenchimento deste campo pode ser combinado com o campo **Classe morfosintática** para refinar a pesquisa.

Pesquisa no corpus

CCP Query: Pesquisar construção da pesquisa | ver | opções

Construção de pesquisa

Pesquisa do texto	Pesquisa do documento
Forma do aluno: igual a	Informante: [selecionar]
Forma normalizada: igual a	Tarefa: [selecionar]
Classe morfosintática: construção de etiquetas	Proficiência: [selecionar]
Lema: termina em	Nacionalidade: [selecionar]
	Língua materna: [selecionar]
	Outra pesquisa
	Tarefa 2 (subtarefas): [selecionar]
	Formal / Informal: [selecionar]
	Ato ilocutório: [selecionar]
	Search within: Text

Pesquisar cancelar ajuda

Pesquisa no corpus

CCP Query: [pos="A.*" & lemma="*.ver"] within text Pesquisar construção da pesquisa | ver | opções

Visualização da pesquisa CQL

1	Classe morfosintática = [starts with] A and Lema = [ends in] ver
---	--

21 resultados

Texto: Transcrição Forma do aluno

Etiquetas: Classe morfosintática Lema

contexto	policia aparece ele fica muito / Ah //	Agradável	/ ver / o policia // [.] Ah / Um // e / Ah / o
contexto	Um // @ // Isso não é @	aceitável	eu sei // desculpa // Desculpa desculpa
contexto	meu curso e seria muito @	agradável	para // o senhor pode Ah assistir
contexto	está Ah // Ela está muito Ah // Oh // muito	agradável	muita // Ah não sei Ah // E Ah / ela
contexto	também / Ah // Também ficam muito / Ah // Ah // muito	agradável	ver / Ah / o ladrão // E também
contexto	obrigada / gosto muito / é muito /	amável	// Desculpa / senhora professora foi uma
contexto	asi que a educação tá	disponível	para todos // Parabéns [Eva] Ah que
contexto	sei acho que não é	impossível	existe algum coisa urgente / Um // muito
contexto	Com trinta minutos talvez seja	improvável	mas / senhor professor parabéns // não
contexto	não podia chegar @ antes // Muito	possível	? => o que posso fazer
contexto	tenho // Um // a carta // como é	possível	porque eu vou já entregar
contexto	podia escrever [...] o mais breve	possível	tão rapidamente // Não faz mal
contexto	escrever / aquela carta / quando é	possível	estragaste os meus calças favoritas
contexto	sei // // Como é que é	possível	ter => [] [] / => copo // de água geral
contexto	senhor Ah // => Ah // me @	possível	Um fazer munto rápida? / por
contexto	preciso / da carta es	possível	Um escrever a carta de recomendação
contexto	comigo? // // O senhor // seria // @	possível	Um repararos // Muito obrigado // Ah eso me
contexto	zapatos para ver // se é	possível	porque // tenho urgência // Porque tu
contexto	escreva a carta assim que	possível	que / a policia // seja @ amigo
contexto	e ela disse // Como é	terrível	desculpa / // mas Ah tinha um problema
contexto	Desculpa peço desculpa porque foi		

9. Como pesquisar sequências de palavras de acordo com a ordem que ocupam na frase, combinando *Classe morfofossintática / Lema / Forma do aluno*?

- ✍ Esta é uma forma de pesquisa avançada que terá que ser feita preenchendo *manualmente* o campo de pesquisa geral.
- ✍ A pesquisa de um dado segmento obedece à seguinte estrutura:

[etiqueta="classe morfofossintática.*"]
[etiqueta="lema/forma do aluno"]

- ✍ Os códigos para as etiquetas são os seguintes:

pos = classe morfofossintática
lemma = lema
form = forma do aluno
word = forma corrigida

- ✍ As abreviaturas associadas às diferentes classes morfofossintáticas podem ser consultados *aqui* (PDF).

- ✍ Exemplos de pesquisa:

1. [pos="AQ.*"] [pos="N.*"]
(adjetivo qualificativo seguido de nome)

Pesquisa no corpus

COP Query: [pos="AQ.*"] [pos="N.*"] construção da pesquisa | ver | opções

Visualização da pesquisa CCL

1	Classe morfosintática = [starts with] AQ	2	Classe morfosintática = [starts with] N
---	--	---	---

120 resultados • A mostrar 100 - 120

Texto:

Etiquetas:

contexto	a fazer? Também ^{pp} muito	obrigado professor	// Muito obrigado // Ainda é mais
contexto	pode ir à minha	pequena festa	com -> os meus pais a
contexto	curso penso ter uma	pequena festa	e como gostei muito das
contexto	gostaria de vir a uma	pequena festa	que vamos a dar nosso
contexto	favor Professora nós ^{pp} tínhamos uma	pequeno festa	na minha casa [.] -> @ => ! ?
contexto	João, um jovem	português O	João mora no Porto
contexto	polícias e E // na	próxima dia	// O ladrão ^{ab} ^{liber} a polícia e
contexto	Professor <-> eu tenho um festa	próxima semana	fim de semana sei você
contexto	cinema comigo na	próxima semana	? Professora ainda não escreveu
contexto	cedo? Não faz isso	próxima vez	// Muito obrigado por avisar-me
contexto	professor ^{pp} // Bom trabalho -> @ no	próximo ano	vou ganhar esse prêmio também
contexto	está passando // E pra o	próximo dia	// O ladrão é -> a mesma
contexto	cinema comigo no	próximo domingo	? -> // Desculpe mas eu tenho
contexto	pão padieiro roupa nuvem leão	selvagem leite	cadeira jogo jogador fado fadista
contexto	cadeira jogo jogador fado fadista	sozinho aquecedor	aquecer
contexto	ou // Acho que -> os homes @	sãos amigos	todos E a mulher ^{ab} ^{liber} não
contexto	terra carro estafa		
contexto	D	velha mar	vaga sonho doce queixo bela
contexto	cola		
contexto	Encontrou ^{ab} que encontrou / @ ^{ab} as / ^{liber} // A	velho amiga	dela e mulher ^{ab} / ^{angry}
contexto	que o juiz também é	velho amigo	dos dois delinquentes então
contexto	ver o polícia ^{eran} dois	velhos amigos	abrazam-se e a senhora

2.

[pos="SP.*"] [lemma="casa"]

(nome "casa" antecedido de preposição)

Pesquisa no corpus

COP Query: [pos="SP.*"] [lemma="casa"] construção da pesquisa | ver | opções

Visualização da pesquisa CCL

1	Classe morfosintática = [starts with] SP	2	Lema = casa
---	--	---	-------------

2 resultados

Texto:

Etiquetas:

contexto	problema com as minhas companheiras de casa ^{pp} desculpa // Fico feliz e por esse
contexto	dia um // Uma senhora ^{ab} volta para casa mas um homem -> que tem

Descarregar resultados - Memorizar expressão de busca

Opções de frequência

Colocação por: [selecionar] | Tamanho do contexto: [1] | Direção: [Esquerda e direita] |

Frequência por: [selecionar]

3.

[form="se" & pos="PP.*"] [pos="V.*"]

(clítico+verbo)

É possível introduzir uma restrição à **forma do aluno** para obter apenas resultados para a estrutura pretendida (clítico+verbo).

Pesquisa no corpus

CQP Query: [inform="se" & pos="PP-*"] [pos="V-*"] construção da pesquisa | ver | opções

Visualização da pesquisa CQL

1	2
Forma normalizada = se	Classe morfosintática = [starts with] PP
	Classe morfosintática = [starts with] V

60 resultados

Texto:

Etiquetas:

contexto	estavam muito felizes // de verles /	se abraçaram	E a senhora se foi
contexto	su amigo E os dois	se abraçaram	a senhora / foi muito zangada
contexto	de polícia não sei como	se chama	guarda para dizer que este
contexto	O senhor diretor @ Ah queremos @ @ [] como	se chama	convidar / o senhor pelo
contexto	o / Eu não sei como	se chama	essa // O juiz / Juiz entra
contexto	entra o // Não sei como	se chama	[] // este homem que // o / // juiz /
contexto	também levanta os mãos e	se començam	/ os / homens // Abraçar? não
contexto	o ladrão [] // Se parece que	se conhecem	de toda a vida // abraçam-se
contexto	eles se deram conta que	se conhecem	a mulher se vai // de
contexto	juiz? // E os homens	se conhecem	também
contexto	Caso / // no recuerdo [] // E também	se conhecem	todos e / // se parece muito
contexto	polícia levanta os mãos que	se conhecem	// e / // Não sei / enfim são
contexto	verdade e aquele ladrão @	se conheceram	/ não se conheciam // [] / senhora
contexto	o ladrão e a / polícia // @	se conheceram	// E se vieram / e / também
contexto	aquele ladrão // se conheceram não	se conheciam	// [] / senhora ficou muito surpreendida
contexto	aquele ladrão e a polícia	se conheciam	// E ela achava muito injusta
contexto	juiz // Eles todos eles três	se conheciam	/ e depois ele // // e depois
contexto	verdade // // Estas pessoas	se conheciam	e // pronto // // E ela
contexto	espelado que estas pessoas todas	se conheciam	porque um é uma // E
contexto	Mas na verdade eles	se conheciam	todos e por isso acho
contexto	a polícia e o ladrão	se conheciam	e / E bom que era
contexto	E o / juiz eram amigos	se conheciam	/ então eles também / começaram a

4.

[pos="V.*"] [lemma="se"]

(verbo+clítico)

Pesquisa no corpus

CQP Query: [pos="V.*"] [lemma="se"] construção da pesquisa | ver | opções

Visualização da pesquisa CQL

1	2
Classe morfosintática = [starts with] V	Lema = se

61 resultados

Texto:

Etiquetas:

contexto	acha / que / o problema pode	resolver-se	/ agora // // mas depois o homem
contexto	polícia estão / preocupados // [] @ // tudo vão	resolver-se	porque também o / juiz é
contexto	a sua mala E / // depois /	apresentou-se	=> o polícia // Mas // foi engraçado
contexto	cumprimentaram e ficaram felizes por	encontrar-se	/ e a senhora ficou chateada
contexto	são amigos // // Portanto a // mulher // //	sintou-se	muito // // Muito [] / não não muito
contexto	muito / // // Muito [] / não não muito /	Sintou-se	muito mal não // // não encontro
contexto	o ladrão / e o polícia //	Encontraram-se	e // reconheceram-se / Descobriram que
contexto	a trinta minutos / // a professora // // @ //	reconheceram-se	/ // Descobriram que / eles são bons
contexto	ele // abraçaram-se e também //	importava-se	de agora escrever / ter de
contexto	ladrão @ abraçaram-se e também //	abraçaram-se	e também // cumprimentaram-se e
contexto	depois ele // // e depois eles	cumprimentaram-se	e ficaram todos muito contentes
contexto	temos / eles / // não sei / eles @ //	cumprimentaram-se	e todos ficaram muito felizes
contexto	o seu amigo melhor // // eles	abraçaram-se	// // aqui // Não compreendo / aqui
contexto	o polícia e o ladrão // //	abraçaram-se	/ e senhora estava muito irritada
contexto	mulher fica // // chateada // // // E ela	relembrou-se	que são amigos // e // // Dão
contexto	gabinete / da polícia / e	vai-se	embora / Ao gabinete / da
contexto	comigo? // // Olá doutor doutora	queixa-se	/ E // depois há uma cena
contexto	amigos // // Amigos da infância //	lembrou-se	da carta de recomendação
contexto	conhecem de toda a vida // //	abraçaram-se	// // E a mulher / fica / com
contexto	muito tempo mas / o professor	abraçaram-se	// // A senhora / acha // isto insolito
contexto	E ele o polícia // // deram //	lembra-se	=> // // da carta // de reclamação
contexto		Deram-se	um abraço // E a juíza

10. Como pesquisar estrangeirismos? *Quieres venir?*

- As formas que sofreram algum tipo de transferência da L1, ou de outras L2, encontram-se sombreadas no texto transcrito a lilás.
- A única forma de pesquisar estas formas é selecionar **Foreign text**, na opção **Search within**, na área de pesquisa e clicar em **Pesquisar**.

(*o aspeto gráfico desta opção ainda está em desenvolvimento.)

Pesquisa no corpus

CQP Query: construção da pesquisa | ver | opções

Construção de pesquisa

Pesquisa do texto

Forma do aluno: igual a

Forma normalizada: igual a

Classe morfosintática: construção de etiquetas

Lema: igual a

Pesquisa do documento

Informante: [selecionar] ▼

Tarefa: [selecionar] ▼

Proficiência: [selecionar] ▼

Nacionalidade: [selecionar] ▼

Língua materna: [selecionar] ▼

Outra pesquisa

Tarefa 2 (subtarefas): [selecionar] ▼

Formal / Informal: [selecionar] ▼

Ato ilocutório: [selecionar] ▼

Search within: Foreign text ▼

Text

479 resultados • A mostrar 0 - 100 (seguintes)

Texto:

Etiquetas:

contexto	não muito íntima <i>Ab</i> eu queria <-> <i>Ab</i>	convidar-te <-> <i>Ab</i> pa ir ao
contexto	corrupção <i>nh</i> / mais ou menos assim / <i>Ab</i> / <->	no outro dia aparece # / Todas
contexto	/ chegar <-> / muito atrasado # <i>Ab</i> @ <-> @ / <->	Lamento-me muito / eu atrasado / porque
contexto	Uma moer camina <-> # <i>Ab</i> #	na rua # <i>Uhm</i> # / <i>Uhm</i> /
contexto	# E tira a carteira <-> <-> @ <i>Ab</i> <i>En</i>	tira-la # a carteira / e depois
contexto	com suas @ mãos / altos / Pero <-> @ <i>En</i>	neste momento / vem um policia
contexto	e # E é um amigo <-> @ {	do ladrão / a mulher @ está
contexto	o roubo # [.] # Turned his back	? Turned his face? # <i>Ab</i> o
contexto	back? Turned his face	? # <i>Ab</i> / o policia estava atrás das
contexto	favor # <i>Uhm</i> por favor / passe-me <-> /	? sim o sumo # <i>Ab</i> o senhor
contexto	cinema <-> à noite <-> <->	? quieres venir? # Também muito
contexto	e na terceira @ <i>Va</i> #	Abrraccio não sei como se diz
contexto	Zangada E <i>Ab</i> <i>Ab</i> vai à #	Bureau de police? bureau de
contexto	outros <i>Uhm</i> policia # Então <i>Uhm</i> # na #	? não não é é
contexto	E descontenta não é feliz / <i>En</i> #	El Va ela / <i>En</i> va <-> / ao
contexto	chegar / tão tarde / alguma excusa #	En hora buena no sé <i>En</i> @ meu
contexto	denunciar # O que aconteceu # e #	En nos tribunais # Quando chegou
contexto	policia e o ladrão estão # <i>Uhm</i> # <i>Ab</i> #	En corte para <i>nh</i> # <-> para este assunto
contexto	volta para ver o policia	Eran dois velhos amigos / abrazam-se
contexto	E # ir # [.] # <i>Uhm</i> ir / à #	Estación / police não [.] # Departamento / <i>nh</i> policia # <i>Uhm</i> <i>Uhm</i> # Para
contexto	Desculpa profesora ontem eu desculpa #	I wanted to say yesterday I
contexto	I wanted to say yesterday	I went to but late but
contexto	went to but late but	I do not say I went
contexto	but I do not say	I went <i>nh</i> / eu fui / eu fui / desculpa porque
contexto	rua / e / depois / <i>Uhm</i> / ela é #	I don't really have vocabulary for
contexto	eu digo [.] / pode ser <i>En</i> necessário / <i>Ab</i> [.]	I don't know how to say
contexto	um ladrão <i>Ab</i> # Le # surpreenda # E #	I don't know this word # e
contexto	não pude fazer nada <i>Ab</i> desculpa	I'll make it up to you

11. A pesquisa de dados de natureza pragmática é possível? Parabéns!

- ✍ A primeira etapa para pesquisar dados de natureza pragmática é pesquisar por **Tarefa** e selecionar as que se referem a tarefas de produção oral (tarefas 1, 2, 3 e 4).

Pesquisa no corpus

CQP Query: construção da pesquisa | ver | opções

Construção de pesquisa

Pesquisa do texto

Forma do aluno: igual a

Forma normalizada: igual a

Classe morfosintática: construção de etiquetas

Lema: igual a

cancelar | ajuda

Pesquisa do documento

Informante: [selecionar] ▼

Tarefa: [selecionar] ▼

Proficiência: [selecionar] ▼

Nacionalidade: T1 ▼

Língua materna: T2 ▼

T3 ▼

T4 ▼

T5 ▼

T6 ▼

T7 ▼

Outra pesquisa

Tarefa 2 (subtarefas): [selecionar] ▼

Formal / Informal: [selecionar] ▼

Ato ilocutório: [selecionar] ▼

Search within: Text ▼

- ✍ No caso da **Tarefa 2 - Elicitação de atos ilocutórios** - é possível pesquisar em função de dois critérios, selecionando as opções disponíveis.

i. Situação Formal/Informal

ii. Ato ilocutório

Nota: Este campo ainda se encontra em atualização, uma vez que o processo de anotação textual ainda não está concluído.

Outra pesquisa

Tarefa 2 (subtarefas) [selecionar] ▼

Formal / Informal [selecionar] ▼

Ato ilocutório [selecionar] ▼

- ✍ Pode conjugar este tipo de pesquisa com qualquer outro campo de pesquisa e obter dados de natureza pragmática. Eis alguns exemplos de combinações:

a. Com a opção **Lema**, preenchido com uma expressão de felicitações (por exemplo, “Parabéns”).

Pesquisa no corpus

QOP Query: [lemma = "parabéns"] .. match_text_task = "T2" within text [Pesquisar] construção da pesquisa | ver | opções

Visualização da pesquisa OCL: [parabéns]

Lema = parabéns Tarefa = T2

78 resultados

Texto: [Transcrição] [Forma do aluno]

Etiquetas: [Classe morfosintática] [Lema]

contexto não pude telefona-te desculpa // | Parabéns // e // @ // bom futuro e / bom
contexto eu não atrasar antes desculpa // | Parabéns // | Parabéns
contexto não atrasar antes desculpa // | Parabéns // | Parabéns
contexto um resultado / força / boa sorte // Parabéns muitas felicidades pelo teu
contexto Sim também parabéns muitos parabéns // Parabéns fazes muito bem
contexto corre bem / consigo / muitos parabéns // Parabéns muitos parabéns e fizeste muito
contexto Não tenho ideia / não sei // Parabéns então muitos parabéns // não sei tu
contexto Desculpa // | Desculpa // Os meus parabéns // Parabéns
contexto sinto muito pela / tardança // Parabéns não sei // Parabéns estou muito
contexto / tardança // Parabéns não sei // Parabéns estou muito orgulhosa
contexto educação tá disponível para todos // Parabéns // [Eva] // que bem que tu
contexto e / bom trabalho para todos // Parabéns
contexto Desculpa muito / desculpa por atrasado [.] // Parabéns também [.] eu disse parabéns // e
contexto ganhou a // // // nomeado de diretora // Parabéns // é natural tu ganhaste // este
contexto vocabulário // é fraco // Não parabéns talvez // Parabéns
contexto / tardança // / Desculpa pela tardança // Parabéns // Parabéns
contexto / tardança // / Desculpa pela tardança // Parabéns // Parabéns
contexto melhores feitas com sua direção // Parabéns // // este é um gran momento
contexto Eu diria o mesmo parabéns // Parabéns
contexto não pude // // apanhar o comboio // Parabéns // Parabéns
contexto pude @ // // apanhar o comboio // Parabéns // Parabéns

Obter a frequência destes resultados por **nacionalidade**.

Distribuição no corpus

Expressão de busca: Lema = parabéns

Tarefa = T2

Agrupamento de pesquisas: Nacionalidade

Gráfico: [Tabela] | Contagem: [Contagem] | Guardar: [selecionar]

Grupo	Contagem
Venezuelana	4
Timorense	2
Sul-coreana	2
Polaca	2
Norueguesa	1
Mexicana	4
Luso-americana	2
Japonesa	5
Italiana	4
Indiana	6
Holandesa/ Americana	1
Holandesa	1
Francesa	2
Espanhola	2
Chinesa	18
Britânica	5
Bielorrussa	2
Americana	13
Alemã	2

Ajuda • URL direto

b. Com **Verbos** que denotam atos de fala.

Pesquisa no corpus

CQP Query: `[pos = "Fr.*"] :: match text_task = "T2" within text` construção da pesquisa | ver | opções

Visualização da pesquisa CQL:

171 resultados • A mostrar 0 - 100 (seguintes)

Texto:

Etiquetas:

contexto	cinema comigo hoje ?	por favor //	Boa tarde /	desculpe
contexto	a carta como é possível ?	->	o que posso fazer agora	
contexto	o que posso fazer agora ?	pra mim é muito importante		
contexto	Podes ir buscar @ meu sumo ?	//	Gostaria de convidar //	o senhor
contexto	festa //	Vamos a cinema hoje ? //	Senhor professor eu preciso muito	
contexto	mim O que tu fizeste ?	@@	não posso usar	ok ok
contexto	um copo de água ?	ou <= //	faz favor não boa	
contexto	um copo de água ?	estou com sede //	Passa cá	
contexto	ir comigo ao cinema ?	//	o professor porque não tens	
contexto	sumo de laranja por favor ?	@@	sim? //	boa tarde é
contexto	laranja por favor?	sim ? //	boa tarde é
contexto	a filme à noite ?	//	Obrigada eu tenho estudar muito	
contexto	dar @ o suma por favor ?	//	Desculpa professora	esta noite é
contexto	ir lá na festa ?	//	Olá //	queres ir ao
contexto	queres ir ao cinema ?	//	<=>	Obrigada para me ajudar //
contexto	diar @ @ //	um copo de água ?	//	ou depois podia dar-me
contexto	podia dar-me alguma água ?	ou então ajuda //	Ei pá	
contexto	podes passar-me @ o sumo ?	@@	se faz favor //	então cara
contexto	minha /	cerimónia //	ou qualquer coisa ?	Sim //
contexto	curso /	por favor ? //	Queres ir a cinema	comigo
contexto	Queres ir a cinema /	comigo ? //	O professor peço imensa desculpa	
contexto	pa fazer isto	imdiaalmente ? //	Porquê	fezes isto? //

e. Com Interjeição.

Pesquisa no corpus

CQP Query: `[pos = "Fr.*"] :: match text_task = "T2" within text` construção da pesquisa | ver | opções

Visualização da pesquisa CQL:

78 resultados

Texto:

Etiquetas:

contexto	atenta no meu trabalho //	Ai	obrigada querida eu gosto muito
contexto	ser muito úteis muito obrigado //	Ai	/ muito lindo //
contexto	grave e @ tive de resolver //	Ai	desculpa desculpa eu sinto muito
contexto	a trinta minutos //	Eh	pá leva - //
contexto	água? ou então ajuda //	Ei	as minhas calças
contexto	filme à noite? //	Obrigada	pá podes passar-me @ o
contexto	ir ao cinema? //	Obrigada	eu tenho estudar muito para
contexto	@@	Obrigada	para me ajudar //
contexto	@@	Obrigada	<=>
contexto	@@	Obrigada	tão bonito
contexto	@@	Obrigada	querida //
contexto	@@	Obrigada	Desculpe //
contexto	@@	Obrigada	Desculpa //
contexto	@@	Obrigada	Os meus
contexto	@@	Obrigada	//
contexto	@@	Obrigada	Obrigadíssima //
contexto	@@	Obrigada	Desculpa pela
contexto	@@	Obrigada	terdância
contexto	@@	Obrigada	ou / obrigada professora //
contexto	@@	Obrigada	é muito
contexto	@@	Obrigada	é muito querida gosto muito
contexto	@@	Obrigada	muito //
contexto	@@	Obrigada	Desculpa porque estou atrasado
contexto	@@	Obrigada	pela sua ajuda //
contexto	@@	Obrigada	agora
contexto	@@	Obrigada	obrigada pela flor //
contexto	@@	Obrigada	Não
contexto	@@	Obrigada	muito //
contexto	@@	Obrigada	senhor professor obrigada muito
contexto	@@	Obrigada	//
contexto	@@	Obrigada	Desculpa também //
contexto	@@	Obrigada	desculpa
contexto	@@	Obrigadinho	professor / obrigado //
contexto	@@	Obrigado	Muito obrigado //
contexto	@@	Obrigado	Desculpe
contexto	@@	Obrigado	<=>
contexto	@@	Obrigado	lamento que cheguei muito tarde
contexto	@@	Obrigado	//
contexto	@@	Obrigado	Sinto muito pelo atraso
contexto	@@	Obrigado	ajudar-me //
contexto	@@	Obrigado	Muito obrigado eu
contexto	@@	Obrigado	//
contexto	@@	Obrigado	por a sua ajuda /
contexto	@@	Obrigadíssima	Muito
contexto	@@	Obrigadíssima	na mesma coisa desculpa //

12. Como obter a frequência de determinadas ocorrências em função de diferentes critérios? *É possível!*

- 📄 Após a exibição dos resultados de uma pesquisa, percorrer os resultados até ao final da página onde se encontram os campos relativos a **Opções de frequência**.
- 📄 Ao seleccionar a opção pretendida, serão apresentados os resultados em tabela.
- 📄 **Nota:** Presentemente, a opção **Custom distribution** está a ser atualizada pelo que ainda não é possível utilizá-la.

Pesquisa no corpus

CQP Query: [tema = "ser"] [tema = "possível"] within text [Pesquisar] construção da pesquisa | ver | opções

Visualização da pesquisa CQL

1 Lema = ser 2 Lema = possível

7 resultados

Texto: [Transcrição] [Forma do aluno]

Etiquetas: [Classe morfosintática] [Lema]

contexto	senhor	me	preguntava se	era possível	/ ter	copo	de água geral
contexto	cine comigo?	O senhor	seria	possível	escrever	a carta de recomendação	
contexto	não tenho	a carta	como	é possível	o que posso fazer		
contexto	pudesse escrever	aquele	carta	quando	é possível	tão rapidamente	Não faz mal
contexto	[seleccionar]	Como é que		é possível	estragaste os meus calças favoritas		
contexto	Informante	s para ver	se	é possível	repararos	Muito obrigado	eso me
contexto	Tarefa	disse	Como	é possível	que	a policia	seja
contexto	Proficiência						

Descarregar res: [Lingua materna] [expressão de busca]

Tarefa 2 (subtarefas)

Opções de fr: [Formal / Informal]

Alto Jicatório

Colocação por: [Texto] [Custom distribution]

Frequência por: [seleccionar]

no do contexto: [1] | Direção: [Esquerda e direita] | Submeter

Distribuição no corpus

Expressão de busca: Lema = possível

Agrupamento de pesquisas: Proficiência

Gráfico: [Tabela] | Contagem: [Contagem] | Guardar: [seleccionar]

Grupo	Contagem
C1+	2
B2	2
B1	1
A2	2

[Ajuda](#) • [URL direto](#)

13. Quais os aspetos a considerar na pesquisa por tarefa?

📖 Quando pesquisamos por tarefa, devemos ter em conta os seguintes aspetos:

i. Algumas tarefas não foram transcritas por não reunirem as qualidades técnicas pretendidas, por esta razão, alguns informantes não apresentam ficheiros relativos a essas tarefas.

ii. Os alunos de A1 não realizaram os atos de *censura* e *felicitações* referentes à tarefa 2.

iii. Existem tarefas de produção oral (1, 2, 3 e 4) e tarefas de leitura oral (5, 6 e 7). Os estímulos correspondentes a cada uma delas encontram-se na secção *Estímulos*, [aqui](#).

iv. Relativamente à **tarefa 2**:

📖 Os atos ilocutórios estão numerados de 1 a 6 de acordo com a descrição da *Metodologia*, que pode ser consultada [aqui](#).

1. Pedido

2. Convite/Sugestão

3. Censura

4. Agradecimento

5. Pedido de desculpas

6. Elogio/Felicitações

Tarefa 2

Elicitação de atos ilocutórios

Ato ilocutório	Situação simulada - estímulo	
	Contexto informal	Contexto formal
<i>pedido</i>	<i>Imagine que está a jantar com a sua família e quer que o seu irmão lhe passe o sumo que está longe de si. O que é que lhe dizia / diria?</i>	<i>Imagine que está com sede e entra num café para pedir uma água ao empregado. O que é que lhe dizia / diria?</i>
<i>convite</i> <i>sugestão</i>	<i>Imagine que quer convidar uma amiga para ir ao cinema consigo. O que é que lhe diria?</i>	<i>Imagine que acabou o seu curso e vai dar uma festa. Gostaria de convidar o seu professor favorito. O que é que lhe dizia / diria?</i>
<i>censura</i> *)	<i>Imagine que o seu irmão estragou as suas calças preferidas. O que é que lhe dizia / diria?</i>	<i>Imagine que um dos seus professores prometeu escrever-lhe uma carta de recomendação. Quando a vai buscar, ela ainda não está escrita e a sua entrevista de emprego tem lugar daí a 30 minutos. O que é que lhe dizia / diria?</i>
<i>agradecimento</i>	<i>Imagine que a sua maior amiga lhe oferece um vaso com a sua flor preferida. O que é que lhe diria?</i>	<i>Imagine que um dos seus professores lhe dá instruções muito úteis para a realização de um trabalho. O que é que lhe dizia / diria?</i>
<i>pedido de desculpas</i>	<i>Imagine que tinha marcado um encontro com uma colega, mas chega 30 minutos atrasado. O que é que lhe dizia / diria?</i>	<i>Imagine que tinha marcado um encontro com uma professora, mas chega 30 minutos atrasado. O que é que lhe dizia / diria?</i>
<i>elogio</i> <i>felicitações</i> *)	<i>Imagine que a sua melhor amiga recebeu o prémio de melhor aluna do ano. O que é que lhe dizia / diria?</i>	<i>Imagine que o seu professor preferido foi eleito reitor da universidade. O que é que lhe dizia / diria, se o encontrasse?</i>

🔗 É possível pesquisar por subtarefas, associando o Ato ilocutório à Situação.

Nota: Neste caso, a pesquisa deve ser feita selecionando a subtarefa pretendida, tendo em conta que a letra **F** refere-se a *formal* e a letra **I** a *informal*.

Pesquisa no corpus

CQP Query: construção da pesquisa | ver | opções

Construção de pesquisa

Pesquisa do texto

Forma do aluno: igual a

Forma normalizada: igual a

Classe morfosintática: construção de etiquetas

Lema: igual a

Pesquisa do documento

Informante: [selecionar]

Tarefa: [selecionar]

Proficiência: [selecionar]

Nacionalidade: [selecionar]

Língua materna: [selecionar]

Outra pesquisa

Tarefa 2 (subtarefas): [selecionar]

Formal / Informal: [selecionar]

Ato ilocutório: [selecionar]

Search within: Text

formato TEI (Text)

Os textos que constituem o Corpus Oral de Português L2 - Coimbra (COraL-Co) estão armazenados num banco de dados (CQL - Corpus WorkBench Query Encoding Initiative) para poderem ser pesquisáveis online através do sistema CQL (Corpus WorkBench Query Encoding Initiative).

Para fazer uma pesquisa no corpus de PLE, preencha os campos disponíveis na construção de pesquisa.

🔗 Para que a visualização dos resultados seja mais clara, deve pesquisar por **Utterance search**, em **Opções de de busca**.

Pesquisa no corpus

CQP Query: construção da pesquisa | ver | opções

Opções de busca

Tipo de representação visual: KWIC Context

Mostrar contexto: Tokens: 30 Utterance Search

Ordenar por: Palavra

Estratégia de combinação: Combinação mais longa

Construção de pesquisa

Pesquisa do texto

Forma do aluno: igual a

Forma normalizada: igual a

Classe morfosintática: construção de etiquetas

Lema: igual a

Pesquisa do documento

Informante: [selecionar]

Tarefa: [selecionar]

Proficiência: [selecionar]

Nacionalidade: [selecionar]

Língua materna: [selecionar]

Outra pesquisa

Tarefa 2 (subtarefas): F3

Formal / Informal: [selecionar]

Ato ilocutório: [selecionar]

Search within: Text

14. Em que formatos podem ser descarregados os ficheiros?

O COral-Co apresenta dois tipos de textos que podem ser descarregados em formatos diferentes.

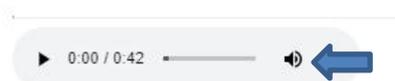
A. PRODUÇÕES ORAIS DOS INFORMANTES

As produções orais podem ser descarregadas em formato **wav** das seguintes formas:

i. Aceda à produção do informante:

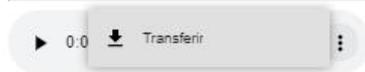
clique do lado direito na **barra de comandos áudio**;

seguidamente, clique em **transferir**.



Opções de representação

Texto: **Transcrição** | Forma do aluno - Mostrar: **Cores** - Etiquetas: **Classe morfosintática** | **Lema** | **Áudio**



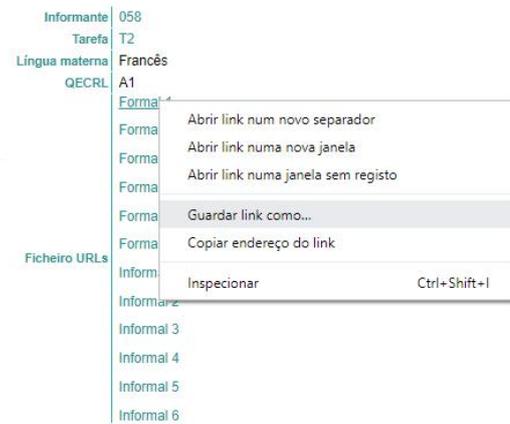
F1 - [Queria] Queria um copo de água <por favor> se faz favor //
I1 - Podes dar-me o sal? sim é o sal? <o> queria o sal se faz favor //
F2 - O professor queria **venir** / à festa? //
I2 - Eu vou ao cinema <esta> à noite <queres> <queres> **queres** **venir**? **quieres** **venir**? //
F4 - Também muito obrigada **nhh** //
I4 - Muito **obrigada** //
F5 - Eu só **n@** não desculpe para o atrasado **senho@** senhor [o] [o] **p@** o professor //
I5 - Desculpe e desculpe para / o atrasado //

ii. Aceda ao cabeçalho da produção:

clique em cima da tarefa;

depois descarregue e/ou guarde o ficheiro áudio correspondente à tarefa.

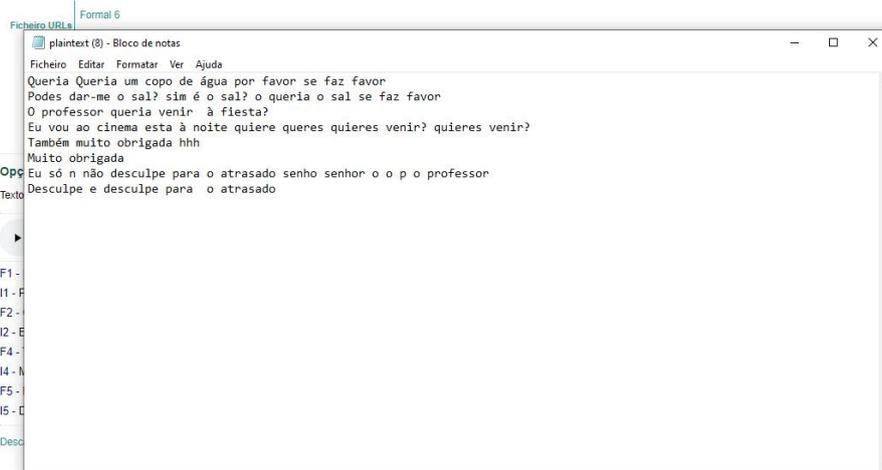
Nota: Consoante o *browser* utilizado, o procedimento poderá ser ligeiramente diferente (ouvir primeiro e transferir a seguir o ficheiro, pedir para guardar primeiro e ouvir depois, clicar com o botão do lado direito do rato).



B. TEXTOS ESCRITOS RESULTANTES DA TRANSCRIÇÃO

- 📄 Os textos escritos podem ser descarregados em formato *txt*.
- 📄 Ao consultar um texto, escolher a **Opção de representação** que pretende obter (*Transcrição*, *Forma do aluno*, *Classe morfosintática* ou *Lema*).
- 📄 Seguidamente, clicar em **Descarregar texto**, no final da página.





15. Como aceder ao perfil dos informantes?

Existem duas formas de aceder ao perfil dos informantes:

ii. consultar, individualmente, o cabeçalho (reduzido) de cada uma das produções escritas;

001_B1_T5	
001_B1_T5	
Informante	001
Tarefa	T5
Língua materna	Italiano
QECRL	B1
Ficheiro URLs	Task 5

iii. consultar os dados dos informantes que se encontram disponíveis na secção **Informantes**, acessível a partir do menu do *corpus*.

Informantes					
	Identificador	Nível QECRL	Género	Língua materna	Nacionalidade
ver	001	B1	F	Italiano	Italiana
ver	003	B1	M	Italiano	Italiana
ver	004	B1	M	Polaco	Polaca
ver	005	B1	F	Lituano	Lituana
ver	006	B1	M	Espanhol	Espanhola
ver	007	A1	F	Lituano	Lituana
ver	008	A1	F	Neerlandês	Belga
ver	009	C1+	F	Chinês	Chinesa
ver	010	C1+	M	Chinês	Chinesa
ver	011	C1+	F	Chinês	Chinesa
ver	012	C1+	M	Japonês	Japonesa
ver	013	C1+	M	Mancanhe	Senegalesa
ver	014	C1+	F	Bielorosso/ Russo	Bielorrussa
ver	015	C1+	F	Alemão	Alemã
ver	016	A1	M	Árabe	Síria
ver	017	A1	M	Finlandês	Finlandesa
ver	018	B2	M	Espanhol	Americana
ver	019	B2	F	Espanhol	Americana
ver	021	A2	M	Coreano	Sul-coreana
ver	022	B2	M	Neerlandês	Holandesa
ver	023	B2	M	N.R.	Chinesa
ver	024	B2	M	Japonês	Japonesa
ver	025	B2	F	Inglês	Americana

 Pode ordenar os ficheiros, clicando em cima de cada um dos campos, para facilitar a procura.

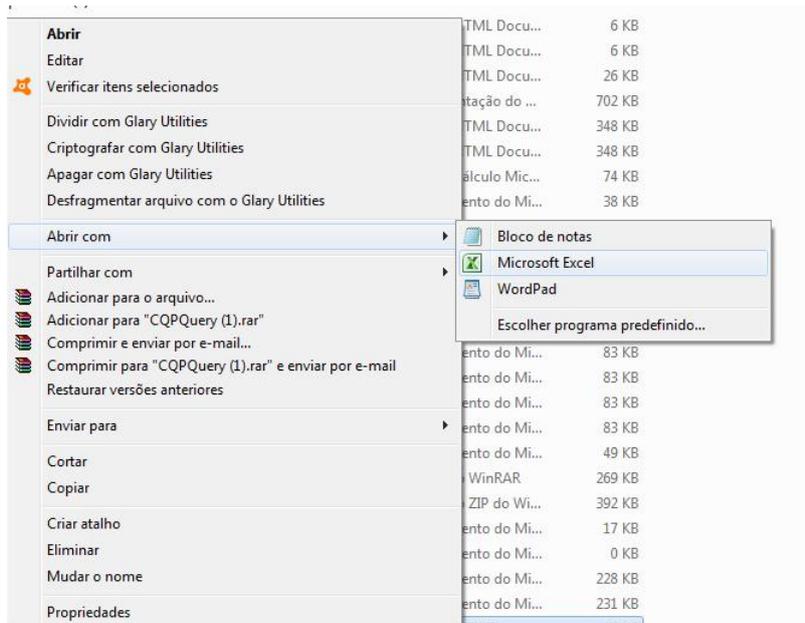
 Clique em **ver** para aceder ao cabeçalho (expandido) com os dados dos informantes.

Informantes

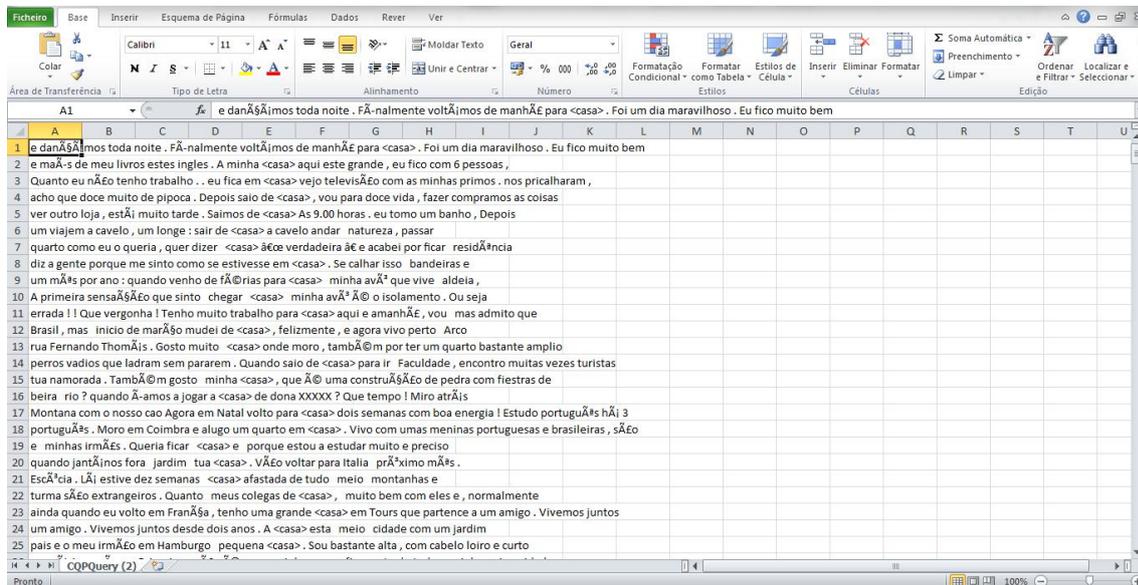
Informante	
Identificador	001
Data de nascimento	1991.12.27
Nível QECL	B1
Fala português fora do contexto escolar?	As minhas companheiras de casa são três meninas portuguesas, os amigos delas e duas minhas colegas com suas amigas.
Género	F
Língua de escolarização	Francês
Língua materna	Italiano
País em que nasceu	Itália
Nacionalidade	Italiana
Países em que já viveu	França/1 mês e meio; Grécia/ 1 semana; Espanha/ 1 semana.
Proficiência em Português	
Produção escrita	B1
Compreensão escrita	B2
Produção oral	B1
Interação oral	B1
Compreensão oral	B2
Outra(s) LNM conhecida(s)	
Outra(s) língua(s) estrangeira(s) conhecidas	Francês; Inglês
LNM em que é mais proficiente?	
Outras línguas não maternas?	Francês
Produção escrita	B1
Compreensão escrita	B2
Produção oral	B1
Interação oral	B1
Compreensão oral	B2

Ficheiros de tarefas para este informante

 **Nota:** Quando acede aos dados dos informantes, também é possível consultar quais as tarefas que cada um deles realizou.



📄 Cada página de resultados apresenta 100 ocorrências de cada vez. Ao descarregar os resultados, basta fazê-lo **uma vez** para guardar, neste exemplo, as 481 ocorrências da palavra “*casa*”, não sendo necessário fazê-lo página a página.



📄 **Nota:** A palavra pesquisada, surge identificada entre < parênteses angulares >.

De seguida, em **Guardar**, seleccionar a opção **CSV**, para transferir os dados .

Distribuição no corpus

Expressão de busca: Lema = amigo
Agrupamento de pesquisas: Língua materna

Gráfico: Tabela | Contagem: Contagem | Guardar: [selecionar] ▼

Grupo	Contagem
Árabe	2
Vietnamita	3
Tétum	1
Polaco	1
Neerlandês	2
N.R.	2
Italiano	4
Inglês	7
Húngaro	1
Espanhol	7
Eslovaco	1
Coreano	3
Concani	1
Chinês	12
Alemão	2

[selecionar] ▼
[selecionar]
SVG
PNG
CSV
JSON

Download data as Comma-Separated Values

[Ajuda](#) • [URL direto](#)

Após a transferência, clicar em **Abrir**.

Distribuição no corpus

Expressão de busca: Lema = amigo

Agrupamento de pesquisas: Língua materna

Gráfico: Tabela | Contagem: Contagem | Guardar: CSV

Grupo	Contagem
Árabe	2
Vietnamita	3
Tétum	1
Polaco	1
Neerlandês	2
N.R.	2
Italiano	4
Inglês	7
Húngaro	1
Espanhol	7
Eslovaco	1
Coreano	3
Concani	1
Chinês	12
Alemão	2

Ajuda • URL



(72)

Finalmente, selecionar o **Excel**, na caixa de diálogo, para abrir o ficheiro .

Expressão de busca: Lema = amigo

Agrupamento de pesquisas: Língua

Gráfico: Tabela

Grupo	Contagem
Árabe	2
Vietnamita	3
Tétum	1
Polaco	1
Neerlandês	2
N.R.	2
Italiano	4
Inglês	7
Húngaro	1
Espanhol	7
Eslovaco	1
Coreano	3
Concani	1
Chinês	12
Alemão	2

Ajuda • URL direto

Abrir com

Escolha o programa que deseja utilizar para abrir este ficheiro:

Ficheiro: transferir (3)

Adobe Acrobat Reader DC Adobe Systems Incorporated	Bloco de notas Microsoft Corporation
Internet Explorer Microsoft Corporation	Microsoft Excel Microsoft Corporation
Microsoft Office 2010 Microsoft Corporation	Microsoft Word Microsoft Corporation
Paint Microsoft Corporation	Visualizador de Fotografias do Windows Microsoft Corporation
Windows Media Center Microsoft Corporation	Windows Media Player Microsoft Corporation
WordPad Microsoft Corporation	

Utilizar sempre o programa seleccionado para abrir este tipo de ficheiro

Procurar...

OK Cancelar

Guardar o ficheiro com o nome pretendido .

📌 Duas opções estão disponíveis:

i. Clicar em **ver** para lembrar os resultados obtidos anteriormente.

📌 Selecionar as pesquisas a comparar e clicar em **Comparar expressões de busca**.

📌 Depois, **Guardar** as expressões de pesquisa em formato **CSV**.

📌 **Nota:** Dependendo do *browser*, as expressões de pesquisa apenas ficam armazenadas temporariamente.



Expressões CQL guardadas

Expressões CQL provisoriamente guardadas

<input type="checkbox"/>	%5Blemma+%3D+%22precisar%22%5D+within+text	editar ver
<input type="checkbox"/>	%5Blemma+%3D+%22importante%22%5D+within+text	editar ver

[Comparar expressões de busca](#)

ii. Clicar em **editar** para preencher/alterar os campos **Nome** e **Descrição**.

📌 Preencha o nome com a palavra/expressão pesquisada.

📌 Finalmente, **Guardar** as expressões de pesquisa de acordo com os dados preenchidos.



Editar expressões CQL guardadas

Nome	<input type="text" value="preciso"/>
Descrição	<input type="text" value="contexto de preciso"/>
Pesquisa CQL	<input type="text" value="[lemma = 'precisar'] within text"/> construção da pesquisa ver

•

📌 Pode, posteriormente, **Comparar expressões de busca** guardadas, selecionando-as.

Expressões CQL guardadas

Expressões CQL provisoriamente guardadas

<input checked="" type="checkbox"/>	preciso	editar	ver	contexto de preciso
<input checked="" type="checkbox"/>	importante	editar	ver	contexto de importante

[Comparar expressões de busca](#)

📄 De seguida, pode guardar esta comparação de pesquisas e voltar a pesquisar as expressões guardadas clicando, primeiro, sobre a palavra/expressão e depois em **Search for**.

📄 **Nota:** Algumas funcionalidades ainda não estão disponíveis nesta área.

Comparação de pesquisas

Gráfico: [Tabela](#) | Guardar: [CSV](#)

Search for Expressão de busca = importante

Expressão de busca	Contagem
preciso	20
importante	5

[Ajuda](#)