Extraction of target structures in learners' corpora: CQL queries for the exploitation of COPLE2

Raquel Amaro[1,2], *Alexandre Carreira*[1,2], *Alice Vieira*[1,2], *Cláudia Castro*[1,2], *Esmeralda Leong*[2]

[1]Centro de Linguística da Universidade Nova de Lisboa
[2]Faculdade de Ciências Sociais e Humanas da Universidade Nova de Lisboa

Abstract

Foreign language (FL) or second language (L2) corpora are sets of productions by non-native speakers, learners of a given language, which contemplate the errors and well-formed structures produced. These serve different research objectives, such as studies on language acquisition (LE and L2), phenomena of linguistic interference or analysis and diagnosis of LE/L2 proficiency levels. In the context of this research, the definition of the learner's proficiency level is often relevant, and this is done, typically, through the analysis of the presence or absence of errors in the learners' productions, based on mappings of typical or expected errors and well-formed structures for a given level of proficiency. However, contrary to the learner's error – which is explicitly marked in the corpus and whose typology and methodology of analysis constitutes a subtopic of investigation on its own –, the well-formed structures, and in particular the target structures (well-formed structures expected in the learners' productions of a given level of proficiency), are not easily identifiable in the corpora. The work presented here aims to fill this gap in COPLE2 – *Corpus of Portuguese Foreign/Second Language* through the use of expressions in CQL – *Corpus Query Language*. Based on pre-identified target structures and on the information made available in COPLE2, such as morphosyntactic tagging and different levels of information and annotation (learner production, teacher correction, normalized form, lemma, etc.), we propose query expressions in CQL that easily allow any user to immediately extract examples of target structures by proficiency level. The construction of the query expressions implies the definition and testing of the best strategies for each case and requires the systematization of linguistic rules and patterns of occurrence of the phenomena in question, but also the definition of ways to circumvent the limitations inherent to the corpus annotation, on the one hand, and the query language, on the other.

Keywords: Corpus Query Language, Learner's Corpus, target structures, PFL, L2.

Resumo

Os *corpora* de língua estrangeira (LE) ou língua segunda (L2) são conjuntos de produções de falantes não nativos, aprendentes de uma dada língua, que naturalmente incluem os erros e os acertos produzidos. Estes *corpora* servem diferentes objetivos de investigação, tais como estudos sobre aquisição de LE e L2, fenómenos de interferência linguística ou análise e diagnóstico de níveis de proficiência de LE/L2. No contexto da investigação destes tópicos, a definição do nível de proficiência do aprendente é muitas vezes relevante e esta é feita, tipicamente, através da análise da presença ou ausência de erros nas produções dos aprendentes, tendo como base mapeamentos entre erros e acertos típicos ou expectáveis para um dado nível de proficiência. No entanto, contrariamente ao erro do aprendente – que é explicitamente marcado no *corpus* e cuja tipologia e metodologia de análise constitui por si um subtópico de investigação –, os acertos, e em particular as estruturas-alvo (estruturas bem formadas expectáveis em produções de aprendentes de um dado nível de proficiência), não são facilmente identificáveis nos *corpora*. O trabalho que aqui se apresenta visa, assim, colmatar essa lacuna no COPLE2 – *Corpus de Português Língua Estrangeira*/*Segunda* através da utilização de expressões de pesquisa em CQL – *Corpus Query Language*. Tendo por base estruturas-alvo pré-identificadas e a informação

disponibilizada no COPLE2, tal como a etiquetagem morfossintática e os diferentes níveis de informação e anotação (produção do aprendente, correção do professor, forma normalizada, lema, etc.), são propostas expressões de pesquisa em CQL que permitem facilmente a qualquer utilizador a extração imediata de exemplos de estruturas-alvo por nível. A construção das expressões de pesquisa implica a definição e a testagem das melhores estratégicas e exige a sistematização de regras linguísticas e de padrões de ocorrência dos fenómenos em causa, mas também a definição de formas de contornar as limitações inerentes à anotação do *corpus*, por um lado, e à linguagem de pesquisa, por outro.

Palavras-chave: *Corpus Query Language*, *Corpus* de aprendentes, estruturas-alvo, PLE, L2.

1. Introduction

Foreign language (FL) or second language (L2) corpora are well-known resources, used for many different research purposes, such as studies in language acquisition, linguistic interference and analysis, diagnosis of FL/L2 proficiency or proficiency profiling (Tracy-Ventura & Paquot, 2020). Depending on the purposes and investment in their compilation and curation, these resources usually contemplate the annotation of the errors produced by the learners. Whatever is not annotated/corrected/identified is a well-formed structure. However, studying FL/L2 phenomena can require as much information on the ill-formed structures as well as on the well-formed structures, at several levels.

For instance, determining the learner's proficiency level requires analysing the presence (or absence) of errors in the learners' productions, which is performed based on the mapping of typical or expected errors, on the one hand, but also of typical or expected well-formed structures for a given level of proficiency (Gramacho et al., 2019; Talhadas, 2016). This means that a B1 level learner is expected not to make common spelling errors, for instance, but also that he/she is expected to produce complex sentences using subordination (and not only simple coordinated ones). However, contrary to the learner's error, which is typically explicitly marked in these corpora and whose analysis and typology constitutes a research topic on its own (Castello et al. 2016), the well-formed structures, i.e., whatever is consistent with the grammar rules of the FL/L2 is not marked. But this does not necessarily mean that all that is not marked, and therefore, well-formed, is of the same nature or relevance. In particular, to perform many of the analysis mentioned earlier, it is necessary to identify target structures, which consist of well-formed structures expected in the learners' productions of a given level of proficiency. However, these productions are not marked, making it quite difficult to extract them from the corpora.

Based on the target structures identified in the POR Nível project (Gramacho et al., 2019) and making use of the information available in COPLE2 – (del Río & Mendes, 2018; Mendes et al., 2016), we present query expressions that allow any user to easily and immediately extract examples of target structures by level. The construction of the query expressions in *Corpus Query Language* (CQL) (Cambridge, 2012; Christ, 1994) implies the definition and testing of strategies and requires the systematization of linguistic rules and patterns of occurrence of the phenomena in question, but also the definition of ways to circumvent the limitations inherent to the corpus annotation, on the one hand, and to the query language, on the other.

In the next sections, we present the process of building the CQL expressions, discussing the methodology used, the list of target structures considered, and the levels and type of information annotated and encoded in COPLE2, as well as the different CQL expressions defined considering the units of analysis or phenomena to be tackled, namely if we were operating at sequences of characters or at word form level, if we were considering multiword expressions or if we were dealing with longer distance phenomena such as subject-verb agreement. We present an analysis and evaluation of the results obtained, focusing on the invalid results extracted, to allow for a more accurate use of the extraction expressions, but also to inform the design of future query expressions. The usefulness and impact of the work developed, as well as possible implementations of the work here

presented, are briefly commented on the final remarks section. All query expressions provided in the paper are ready to be immediately and freely used.

2. Definition of the CQL expressions

The definition of adequate CQL expressions, with good performance, is the main part of the work put forth in this paper. In order to present the reflection and the specific decisions taken in process, we divided this section into several parts: the first consists of a brief presentation of the methodology followed; the second presents the target structures aimed at; the third section describes the information available in COPLE2, which can be used in the extraction queries; the fourth presents and discusses the CQL expressions created; and the fifth and final section evaluates and analyses the results achieved, relating them to the data and resources in use.

2.1. Methodology

To build the CQL expressions for extracting the relevant data (i.e., a specific set of target structures), we devised the following methodology:

1. Identification of the phenomena at stake: selection and description of the relevant target structures, according to existent literature on the subject for European Portuguese.
As further presented in Section 2.2, we collected the target structures from the ones described in the POR Nível project (Gramacho et al., 2019), for the proficiency levels A1 to C1.
However, more than collecting lists of phenomena, it is necessary to investigate and describe how these phenomena reflect in the language data. For instance, target structures related to the use of specific verb tenses and moods can either consist in querying specific part-of-speech tags (e.g., Simple Conditional mood/tense) or specific verb expressions (e.g., Future tense using auxiliary verb *ir* + Infinitive).

2. Mapping of the target structures with linguistic rules and/or occurrence patterns. After establishing the set of target structures to be extracted, we searched for specific examples in the corpus, based on linguistic rules (e.g., the internal structure of the noun phrase in European Portuguese) and or frequent/potentially occurring expressions (e.g., subordinative conjunctions; frequent modal verbs, etc.).

We performed diverse simple queries for each target structure to validate and extend, whenever relevant, the patterns based on linguistic knowledge. This allowed us to identify relevant tag sets and other information included in the retrieved examples (e.g., personal pronouns, common nouns, simple and complex proper nouns), as well as to account for less predictable cases, such as adverbs modifying adjectives in post-adjective position (e.g., "*Pessoas menos fortes* <u>*mentalmente*</u>" ≅ 'People less strong <u>mentally</u>').

3. Building the CQL query expressions. After the mapping phase, building the actual CQL query expression required several other steps to accommodate the limitations of the data available (e.g., COPLE2 does not have syntactic analysis), on the one hand, but also to determine how to use the several layers of annotation available (e.g., normalized form, level of proficiency, tokenization, lemma, part-of-speech tagging).

The main tasks in this phase concerned transforming structural information in linear information (e.g., which elements, in which relative order, and with which level of optionality, compose a Noun Phrase in European Portuguese) and how to use and combine several levels of annotation (document vs. text annotation) in a single query. The queries were built from the simpler to the more complex expressions, or from the more general expressions to the more specific, in an iterative process.

4. Testing and tunning of the CQL expressions. The CQL expressions were tested in the COPLE2 corpus to fine-tune them. This means that queries that over generate (i.e., queries that produced results that were not consistent with the target structures, along with results that were) were complemented with more specific tags or even specific lexical items, to gather gains in precision; queries that under generate (i.e., queries that produced results that were consistent with the target structures but missed many others) were redesigned to be more comprehensive.

5. Evaluation. This final step consists of the overall evaluation of the results achieved with the CQL expressions. This meant using the expressions and analysing, manually, the extracted results. More extensive results were evaluated through the analysis of random samples. The main purpose of this step is to evaluate the productivity of the queries, on the one hand, but also to understand where the extraction errors occurred and why, thus informing the use of the expressions.

This methodology was followed with good results. The following sections present and discuss relevant parts of the steps taken, explaining in detail the options taken.

2.2. **Target structures**

The target structures we considered in the work here presented were the ones considered in the project POR Nível (Gramacho et al., 2019), listed in Amaro et al. (2020). For the purposes of profiling proficiency levels for PLE, the POR Nível project considered the identification of uses divergent and convergent with target structures, having as bases the curricular content of guidance documents, such as PLE Referential Camões (*Referencial Camões PLE*), the syllabuses of the Instituto de Cultura e Língua Portuguesa courses and the syllabuses of the Portuguese Language and Culture Course of the Faculdade de Ciências Sociais e Humanas da NOVA University Lisbon.

The target structures described covered the orthography/spelling domain, morphology and syntax domain and vocabulary. According to the authors, the target structures were identified in specifically compiled corpora for each of the domains. After the identification, the target structures were categorized (Gramacho et al., 2019, p. 174).

Table 1 details the target structures we accounted for, considering the domain (or level of analysis), the phenomenon at stake, the specific target structure and proficiency level it relates to and an example for each structure.

Table 1. Target structures to be extracted, adapted from Amaro et al. (2020, p. 13)

Domain	Phenomenon	Target structure / Proficiency level	Example
Phonology /Spelling	Grapheme-phoneme correspondence	<ç> with [s] value / A2	*peço* ('I ask')
		<x> with [z] value / B2	*e**x**ijo* ('I demand')
	Stress marking	C1	*vigil**â**ncia* ('survaillance'), *respons**á**vel* ('responsible')
	Nasality before consonant	<m> before <p> or <b>; <n> in the rest of the cases / A2	*co**m**plicado* ('complicated'), *e**n**dereço* ('address')
Morphology & Syntax	Verb Tense, Mood and Aspect	Future with Auxiliary *ir* / A2	***vou morar*** ('I **will live**')
		Simple conditional / C1	***gostaria** de exprimir* ('I **would** like to express')
		Subjunctive *Pretérito Perfeito Composto* / C1	***tenha feito** muitos progressos* ('**had made** many progresses')
	Subject-Verb Agreement	B2	***alguém conhece*** ('**someone knows**')
	Conjunctions and Connectors	*mas* / A1	*Sou suíça **mas** moro em Lisboa há muitos anos* ('I'm Swiss **but** I live in Lisbon since many years now')
		por isso / A2	***por isso** tenho de ir* ('**therefore** I have to go')
		para que / B1	***para que** seja* ('**so** it be')
Lexicon	Specific vocabulary	Academic life / A1	***disciplina** favorita* ('favourite **subject**')
		Urban space / B2	***arredores** das cidades* ('city's **outskirts**')

The thirteen target structures were retrieved from the table presented in Amaro et al. (2020, p. 13), except for i) argument-marker prepositions marking argument of predicative nouns, ii) false friends and iii) idiomatic/frozen expressions. The last two cases were not retrieved because these were not listed in the table. However, if defined as pre-established lists of words or expressions, these can also be easily extracted, as we will demonstrate in the next sections.

2.3. **Information annotated in COPLE2**

The COPLE2 – *Learner Corpus of Portuguese L2* (del Río & Mendes, 2018; Mendes et al., 2016), a resource developed by CLUL – Centro de Linguística da Universidade de Lisboa, is composed of written and oral texts produced by foreign students who are learning Portuguese as a foreign language (PFL) or second language (L2). It also contains language proficiency certification exams. COPLE2 is freely available online (in http://teitok.clul.ul.pt/cople2/) and presents detailed information at the metadata level, concerning the informant and the text produced, as well as various types of linguistic annotation (del Río & Mendes, 2018), as schematized in the Figures 1 and 2 below.

Figure 1. Information in COPLE2 at document level

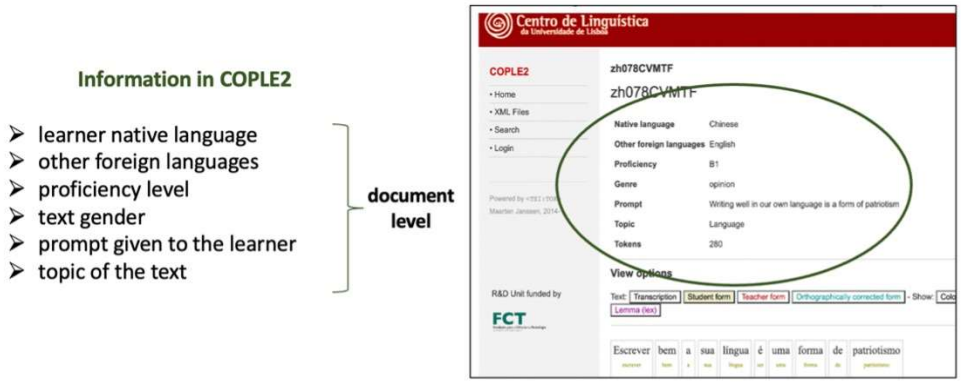
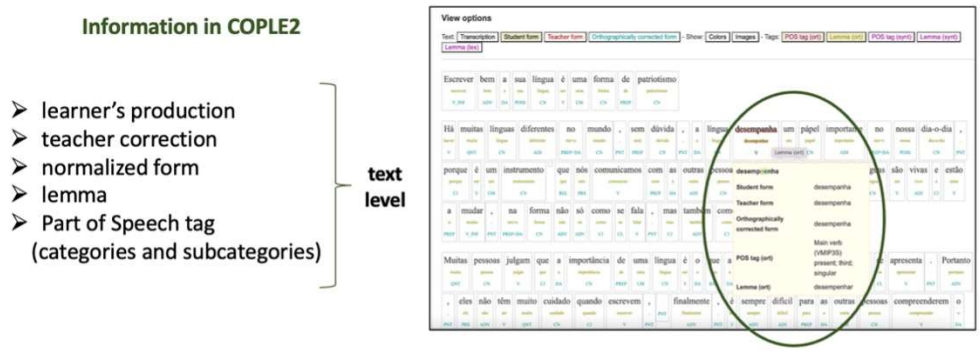


Figure 2. Information in COPLE2 at text level

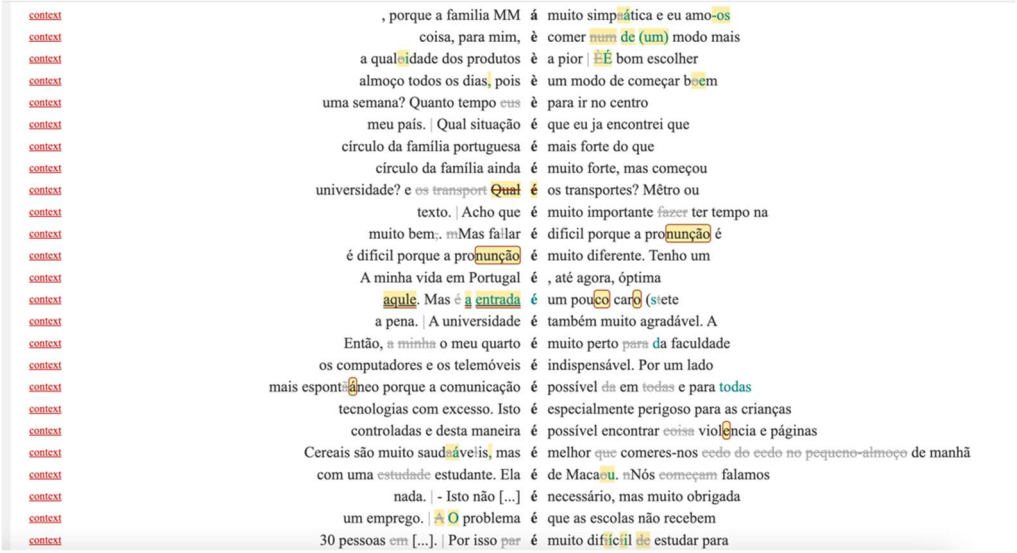


The part-of-speech and lemma information are automatically tagged. The rest of the information is derived from the source texts' metadata (e.g., proficiency level, prompt, topic, learner's native language), retrieved from the source text (teacher's correction), or introduced by the COPLE2 annotators (e.g., normalized form: orthographic normalization, morphosyntactic normalization and lexical normalization).

This rich annotation is associated with a CQL query system, making it possible to combine the different types of features and variables available. The corpus provides data necessary to many research topics, such as such as identifying general errors in PFL/L2 learning or identifying specific errors that may result from transfers of native languages or of other previously acquired foreign languages, enabling the development of applications and didactic materials in the area of PFL/L2 learning and teaching.

The extraction of information requires formulating CQL expressions that can be more or less straightforward depending on the phenomena to be extracted. If these correspond to phenomena specifically annotated (and tagged) in the corpus, it is a matter of knowing and using the specific tags. For instance, a CQL query such as `[form!=nform]` extracts orthographic errors, `[form!=reg]` extracts morphosyntactic errors. Also, the several layers of annotation can be visible in the COPLE2 visualizer (see Figure 3 below).

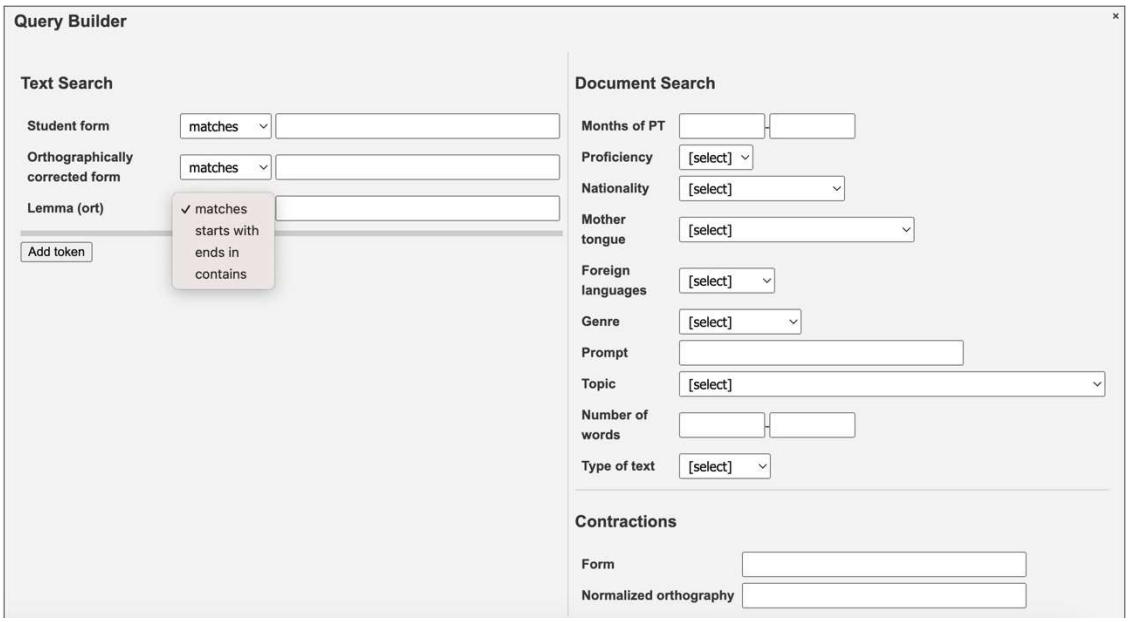
Figure 3. COPLE2 with visualization of error annotation, September 2023



Besides the query command line, the COPLE2 platform provides a Query builder.[1] As shown in the Figure 4 below, this tool helps the user build CQL expressions for one or more token (option 'Add token'), considering three options in the text - 'Student form', 'Orthographically corrected form' and 'lemma' -, and using four string operators - 'matches', 'starts with', 'ends with' and 'contains'. It also allows for combining text and document search parameters (left and right columns of the builder, respectively). This is, in fact, a very useful tool that helped us to know the CQL code for searching parameters at the document level (for instance, 'match.text_mothertongue' is the attribute to search for documents from learners with a specific mother tongue). However, the Query Builder does not replace CQL proficiency, and querying for more complex phenomena requires further knowledge and investment of time.

[1] COPLE2 webpage provides a help page for CQL query builder (http://teitok.clul.ul.pt/cople2/index.php?action=querybuilderhelp) and well as access to further information on CQL in https://cwb.sourceforge.io/documentation.php

Figure 4. COPLE2 Query Builder



The rich annotation system and information used in COPLE2 is a key factor to enable the successful extraction of these structures, as explained in the following sections.

2.4. **CQL expressions**

Considering the target structures to be extracted, and to map the phenomena to specific sets and types of CQL queries, we searched for specific examples in the corpus, based on linguistic rules. This allowed us to identify the relevant tags, possible sets of linear strings of part-of-speech tags, and relevant tags at the document and text levels. From the analysis devised at this stage, it was possible to group the queries by the units these consider - units smaller than words, word forms, multiword expressions, and phrases -, as these require different strategies.

2.4.1. **Sets of characters**

The definition of character sets allowed us to account for the listed phonology/spelling target structures, as these concern the correct use of specific characters in specific sequences, as demonstrated in the Table 2, below.

Table 2. CQL expressions for phonology/spelling target structures

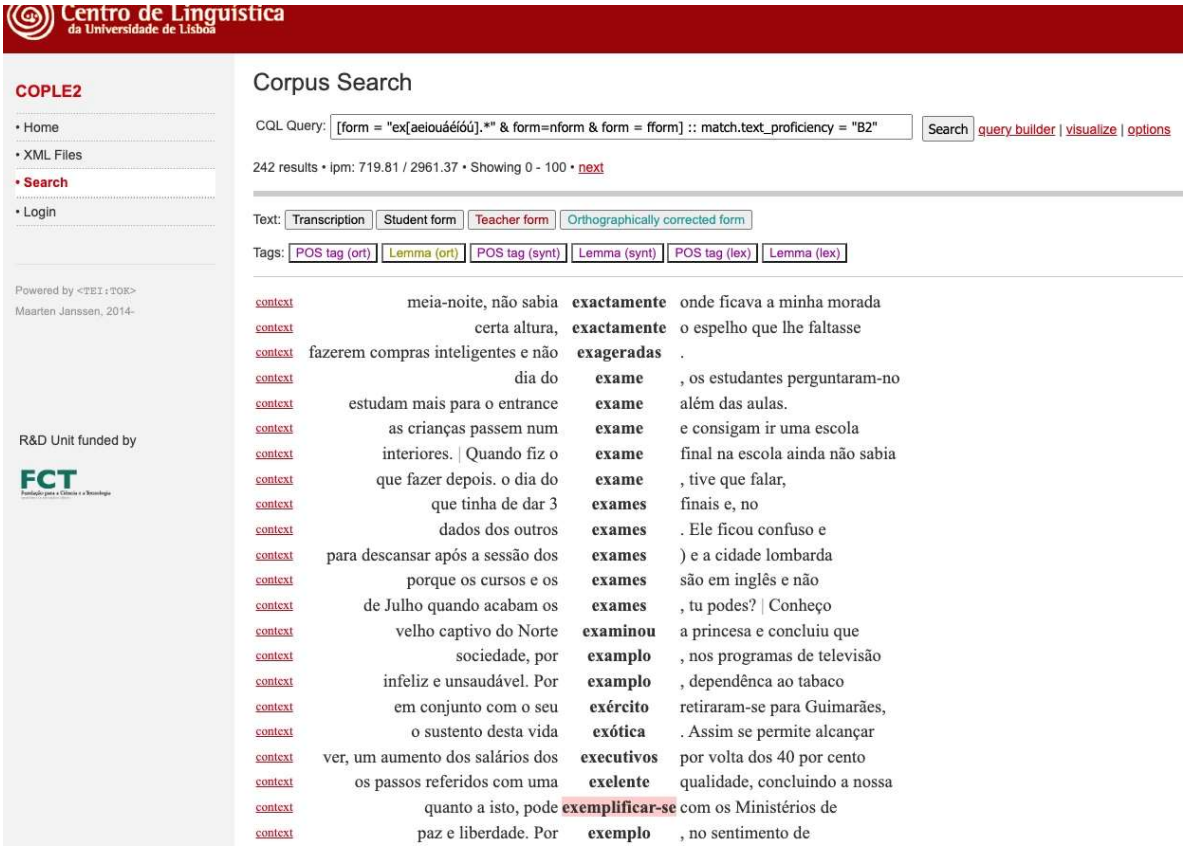
Nr	Target structure / Proficiency level	CQL expression	Results
1	<ç> with [s] value / A2	`[form = ".*ç.*" & form=nform & form = fform] :: match.text_proficiency = "A2"`	594
2	<x> with [z] value / B2	`[form = "ex[aeiouáéíóú].*" & form=nform & form = fform] :: match.text_proficiency = "B2"`	242
3	stress marking / C1	`[form = ".*(À|Á|É|Í|Ó|Ú|Ã|Õ|Â|Ê|Ô|à|á|é|í|ó|ú|â|ê|ô|ã|õ).*" & form = nform & form = fform] :: match.text_proficiency = "C1"`	5313
4	<m> before <p> or <b>; <n> in the rest of the cases / A2	`([form=".*m(b|p).*" & form=nform & nform = fform] | [form=".*n(b|c|d|f|g|h|j|k|l|m|n|p|q|r|s|t|v|x|z|ç).*" & form=nform & form = fform]) :: match.text_proficiency = "A2"`	5763

For instance, to extract the use of the grapheme <ç> (CQL expression 1), we used the 'form' attribute to search for a specific character in a given word (`form = ".*ç.*"`). To assure that the grapheme was correctly used, we checked if the word form used in the learner's production was coincident with the values in the 'nform' and 'fform' attributes (`& form=nform & form = fform`), that is, cases where there was no normalization of the form (`nform`) and/or no correction of the teacher (`forma`). Since this case is expected for A2 proficiency level students, we restricted the query to the document level value "A2" of the attribute 'text_proficiency'. The same strategies were used in the following cases, with the addition of sets of characters that correspond to letters plus diacritics used in European Portuguese (e.g., `áéíóú`) or the sets of letters corresponding the consonants after <n> or <m>.

Also, CQL uses several notations from regular expressions: operators, such as AND (`&`) and OR (`|`), wildcards to represent variables, such as any character (`.`), quantifiers, such as 0 or more times (`*`), or notations such as one of the characters of the set (`[ ]`) and substrings (`( )`). These can operate at different levels, namely character strings (e.g., `.*m(b|p).*`) or attribute level (e.g., `[form = ".*ç.*" & form=nform & form = fform]`). As described above, COPLE2 also allows the combination of annotations at text and document levels (`[form = ".*ç.*" & form=nform & form = fform] :: match.text_proficiency = "A2"`).

The CQL expressions presented in the table above can be directly used (copied and pasted into the query box) in the search system of COPLE2. Figure 5 below shows the results obtained with the CQL expression 2, <x> with [z] value in B2 proficiency level, at current date.

Figure 5. Query results for the CQL expression nr. 2 - <x> with [z] value / B2, September 2023



2.4.2. **Word lists**

The extraction of several of the target structures required CQL expressions considering specific word lists. The definition of these word lists, in turn, could include simple (or atomic) words and multiword expressions, as well as the delineation of the semantic fields to be consider.

As depicted in the Table 3, CQL expressions considering simple words were used to extract target structures related to the domains of morphology and syntax and of lexicon.

Table 3. CQL expressions for target structures with simple word lists, results September 2023

Nr.	Target structure / Proficiency level	CQL expression	Results
5	Simple Conditional / C1	`[pos="VMIC.*" & form=nform & form=fform & form=reg & form=lex] :: match.text_proficiency = "C1"`	55
6	Atomic connectors / A1 and B1	`[form="mas" & form=nform & form=fform & form=lex & pos="C.*"] :: match.text_proficiency = "A1"`	115
7		`[form="mal|embora" & form=nform & form=fform & form=lex & pos="C.*"] [pos!="V.*" & form=nform & form=fform & form=lex]{0,6} [pos="VMS.*" & form=nform & form=fform & form=lex & form=reg]:: match.text_proficiency = "B1"`	6
8	Academic life vocabulary / A1	`[lemma = "estudar|estudante|aluno|escolar|universidade|aula|turma|cantina|professor|tpc|exercício|colega|exame|curso|teste|estojo|caderno|lápis|biblioteca|currículo|faculdade" & form=nform & form=fform & form=lex]:: match.text_proficiency = "A1"`	485
9	Urban space vocabulary / B2	`[lemma = "edifício|habitante|estrada|cidade|passear|trânsito|zona|rio|margem|bairro|loja|redor|centro|aldeia|rua|apartamento|prédio|local|miradouro|supermercado" & form=nform & form=fform & form=lex] :: match.text_proficiency = "B2"`	299

The definition of these query expressions was based on the compilation of word lists from different sources: atomic connectors were collected from the ones described in Gramacho et al. (2019); the vocabulary related to academic life for A1 level were collected from the Portuguese Foreign Language handbooks *Passaporte para Português (A1/A2)* and *Português XXI Nível A1*; the vocabulary related to urban space for B2 level were collected from the Portuguese Foreign Language handbook *Aprender Português (Nível B2)*.

The specific lists of words collected can change, whenever relevant. For instance, if, in a given class, the vocabulary taught at level A1 focuses on daily life and holidays instead of academic life, the set of lemmas used in the CQL expression above can be replaced. What is relevant here is that is more practical and simpler to use one query expression for simple words and another for expressions.

The combination of features to be used in the queries can differ, depending on the phenomena we are extracting. For instance, to extract vocabulary (simple words), the relevant attribute for the query is 'lemma' (so we can extract all inflected forms of a given word, for instance, for the lemma 'habitante' (≅ inhabitant): *habitante* (≅ inhabitant)*, habitantes* (≅ inhabitants)*, habitantezinho* (≅ little inhabitant), ...). COPLE2 has different levels of normalization that can be used to further improve the queries and that are relevant for these cases: *nform*: orthographic normalization, *reg*: morphosyntactic normalization and *lex*: lexical normalization (cf. del Río & Mendes, 2018; Mendes et al., 2016). This way, to assure we are extracting correct uses we also checked if the values of the attributes 'form', 'nform', 'fform' and 'lex' corresponded. The 'reg' attribute was used to assure the correct use of inflected forms, typically in the case of verbs.

The feature relevant for extracting specific verb forms, such the Conditional Mode/Tense is part-of-speech category, verb, (`pos="`**`V`**`MIC.*"`) and subcategories, in the Conditional Mode/Tense (`pos="V`**`MIC`**`.*"`).

The relevant features to extract appropriate occurrences of atomic connectors, on the other hand, can involve the combination of a specific word form with a specific part-of-speech tag (as in `form="mas" &`

`pos="C.*"]`) or the combination of more features, such as a specific word form with a specific part-of-speech tag (`form="mal|embora" & pos="C.*"`) occurring close to, or at a distance of, 5 words of a main verb in the Subjunctive Mode (`{0,6} [pos="VMS.*"]`), with words that are not verbs in between (`[pos!="V.*"]{0,6}`). That is, combining features of the queried word forms with features from co-occurring forms, as illustrated in Figure 6, line 7, below. Also in these cases, to assure we extracted correct uses, the values of the attributes 'form', 'nform', 'fform', 'lex' and 'reg' were checked for all occurring elements.

Figure 6. Query results for the CQL expression nr. 7 - Atomic connectors / A1 and B1, September 2023

2.4.3. **Multiword expressions**

The CQL expressions considering multiword expressions (MWE), that is, a specific sequence of words, were used to account for target structures related to complex verb tenses (auxiliary verbs + main verbs), complex connectors and vocabulary related to academic life and urban space. As described for simple word lists, the different phenomena can require the different combination of features, and different levels of complexity, as showed in Table 4.

Table 4. CQL expressions for target structures with MWE (results in September 2023)

Nr.	Target structure / Proficiency level	CQL expression	Results
10	Future with Auxiliary *ir* / A2	`[lemma="ir" & pos="VMIP.*" & form=nform & form=fform & form=lex & form=reg] [pos="R.*" & form=lemma]? [pos="VMN" & form=lemma & form!="ir"] :: match.text_proficiency = "A2"`	191
11	Subjunctive *Pretérito Perfeito Composto* / C1	`[pos="VASP.*" & lemma="ter" & form=nform & form=fform & form=lex & form=reg] [pos="VMP" & form=nform & form=fform & form=lex & form=reg] :: match.text_proficiency = "C1"`	2
12	Complex connectors / B1	`[form="já|dado|visto|para|antes|depois|logo|sempre|até|desde|ainda|tanto"& form=nform & form=fform & form=lex] [form="que"& form=nform & form=fform & form=lex] [pos!="V.*" & form=nform & form=fform & form=lex]{0,6} [pos="V.*" & form=nform & form=fform & form=lex & form=reg] :: match.text_proficiency = "B1"`	33
13		`([form="todas" & form=nform & form=fform & form=lex] [form="as" & form=nform & form=fform & form=lex] [form="vezes" & form=nform & form=fform & form=lex] [form="que" & form=nform & form=fform & form=lex])|([form="apesar" & form=nform & form=fform & form=lex] [form="de" & form=nform & form=fform & form=lex]) [pos!="V.*" & form=nform & form=fform & form=lex]{0,6} [pos="V.*" & form=nform & form=fform & form=lex & form=reg] :: match.text_proficiency = "B1"`	5
14	Academic life vocabulary / A1	`[lemma="sala" & form=nform & form=fform & form=lex] [form="de" & form=nform & form=fform & form=lex] [lemma="aula" & form=nform & form=fform & form=lex] :: match.text_proficiency = "A1"`	0
15		`[lemma="trabalho" & form=nform & form=fform & form=lex] [form="de" & form=nform & form=fform & form=lex] [form="casa" & form=nform & form=fform & form=lex] :: match.text_proficiency = "A1"`	9
16	Urban space vocabulary/B2	`[form="a" & form=nform & form=fform & form=lex] [form="pé" & form=nform & form=fform & form=lex] :: match.text_proficiency = "B2"`	3
17		`[lemma="junta" & form=nform & form=fform & form=lex] [form="de" & form=nform & form=fform & form=lex] [form="freguesia" & form=nform & form=fform & form=lex] :: match.text_proficiency = "B2"`	0
18		`[lemma="câmara" & form=nform & form=fform & form=lex] [form="municipal" & form=nform & form=fform & form=lex] :: match.text_proficiency = "B2"`	0

The CQL expressions here presented range from simple word sequences (and not lemma) of 2, 3 or more elements such as *a pé* (≅ by foot), *sala de aula* (≅ classroom), *todas as vezes que* (≅ every time that), to sequences of specific part-of-speech category and subcategories and lemma, such as the Subjunctive tense that is composed of the auxiliary verb *ter* (≅ to have) in the Subjunctive Mode, Present Tense (`pos="VASP.*" & lemma="ter"] [pos="VMP"]`) immediately followed by a main verb in Past Participle form (`[pos="VMP"]`).

To account for some flexible MWE included in this set, such as Future with auxiliary verb *ir* (≅ to go), the CQL expression designed also accommodated optional material that can co-occur within the expression such as adverbs (`[lemma="ir" & pos="VMIP.*" & form=nform & form=fform & form=lex & form=reg] **[pos="R.*" & form=lemma]?** [pos="VMN" & form=lemma & form!="ir"] :: match.text_proficiency = "A2"`). The Figure 7, below, show some of the occurrences extracted with more complex examples.

Figure 7. Query results for the expression 13. Complex connectors / B1, September 2023

2.4.4. **Phrases: agreement**

The target structures requiring the levelling of more complex structures concerned subject-verb agreement. The challenge here is that the corpus, although presenting a fine-grained level of part-of-speech annotation, is not syntactically parsed. This way, to assess subject-verb agreement we need to describe, in a linear form, the sequence of word classes that can occur between the core noun of the subject noun phrase (NP) and the verb, and to encode this in CQL. To do so, it is necessary to describe what can follow the core noun within the noun phrase, which lead us to the need to describe the possible configurations of the noun phrase in European Portuguese. For instance, in the sentence (1), below, we have a prepositional phrase and an adverb between the core elements that must agree (in bold).

(1) O **brasão** de Évora ainda **mostra** esse ato heróico de o Geraldo Geraldes (COPLE2, en020CAATI_1)
' The **coat** of arms of Évora still **shows** this heroic act by Geraldo Geraldes '

We considered the following major configurations for the noun phrase in European Portuguese, represented here in simple context-free grammar notation, where NP stands for noun phrase, AP for adjective phrase, AdvP for adverbial phrase, DP for determiner phrase and PP for prepositional phrase. The round brackets indicate optionality:

(2) NP --> Pronoun
NP --> (DP) (AP) Noun (AP) (PP)
DP --> (Quantifier) Determiner (Possessive)
AP --> (AdvP) Adjective (PP)
PP --> Preposition NP
AdvP --> (Adv) Adv
Determiner --> Article
Determiner --> Indefinite
Determiner --> Demonstrative

The configurations in (2) accounts for NPs such as *ele* (≅ he), *Maria* (≅ Mary) or *o saudável incitamento à participação ativa de toda a sociedade civil na resolução dos seus mais urgentes problemas* (≅ the healthy encouragement of the active participation of all civil society in the solution of its most urgent problems).[2]

Besides the elements that are included in the noun phrase, we also accounted for elements that can occur before the verb, as schematized in (3) below:

(3) NP_Subject (Adverb) (Clitic) Verb

The next step is, thus, to transform these schemas in linear expressions in CQL. As demonstrated below, and since there are no structural elements that we can refer to, these can amount to large expressions.

(4) Adverb phrase structure in CQL
```
[pos="R.*" & form=nform & form=fform & form=lex]? [pos="R.*" & form=nform & form=fform & form=lex]
```
which is equivalent to `[pos="R.*" & form=nform & form=fform & form=lex]{1,2}`

However, given that in our case adverbs are always optional, the formulation to be used can be `[pos="R.*" & form=nform & form=fform & form=lex]{0,2}.`

(5) Determiner phrase structure in CQL
```
([form="tod(o|a)s?" & pos="BQ.*" & form=nform & form=fform & form=lex]? [pos="DA.*|BD.*" & !contr & form=nform & form=fform & form=lex] [pos="BP.*" & form=nform & form=fform & form=lex]?) |
```

[2] Example built based on an excerpt of the editorial note of the newspaper *Expresso* of 24/06/2021. Available at https://expresso.pt/expresso/nota-da-direcao/2021-06-24-Nota-editorial-do-Expresso-a-explicacao-para-um-cabecalho-nao-neutral-c18ec16f

```
([form!="tod(o|a)s?" & pos="BQ.*" & !contr & form=nform & form=fform
& form=lex])
```

The determiner phrase includes quantifiers (`pos="BQ.*"`) and uses the negation of the attribute 'contr' to exclude determiners and quantifiers within contracted forms, such as *dos* (≅ of the) or *nalguns* (≅ in some). This expression allows for the following combinations:

(6) o/este/aquele *Noun* (≅ the/this/that *Noun*)
todo o/este/aquele *Noun* (≅ all the/this/that *Noun*)
o/este/aquele meu *Noun* (≅ the/this/that my *Noun*)
todo o/este/aquele meu *Noun* (≅ all the/this/that my *Noun*)
algum *Noun* (≅ some *Noun*)

(7) Adjective phrase structure in CQL

```
[pos="R.*" & form=nform & form=fform & form=lex]{0,2} [pos="A.*" &
form=nform & form=fform & form=lex]
```

(8) Prepositional phrase structure within noun phrases in CQL

```
[pos="S.*" & form="d.*" & form=nform & form=fform & form=lex]
(([form="tod(o|a)s?" & pos="BQ.*" & form=nform & form=fform &
form=lex]? [pos="DA.*|BD.*" & form=nform & form=fform & form=lex]
[pos="BP.*" & form=nform & form=fform & form=lex]?) |
([form!="tod(o|a)s?" & pos="BQ.*" & !contr & form=nform & form=fform
& form=lex]))? ([pos="R.*" & form=nform & form=fform & form=lex]{0,2}
[pos="A.*" & form=nform & form=fform & form=lex])? [pos="E|N.*" &
form=nform & form=fform & form=lex] ([pos="R.*" & form=nform &
form=fform & form=lex]{0,2} [pos="A.*" & form=nform & form=fform &
form=lex])?
```

Prepositional phrases here consider are the ones introduced by the preposition *de* (≅ of), in contracted and non-contracted forms, typically used within the noun phrase, taking as complement a noun phrase. That explains the removal of the condition `!contr` (not contracted) expressed within the determiner phrase. The expression for the prepositional phrase allows extracting cases such as:

(9a) ... é tão diferente **da arquitectura mexicana**... (COPLE2, es010CVMTD)
'... it's so different **from Mexican architecture**... '
(9b) ... embora seja uma **das minhas férias favoritas**, tenho... (COPLE, en022CAMTF_2
'...even though it's one **of my favourite vacations**, I have...'
(9c) ... uma ilha pequena que faz parte **desse grande país**... (COPLE2, de048CVATI)
'... a small island that is part **of that great country**'
(9d) ... só precisas **de algumas camisolas desportivas**, porque a equipa... (COPLE2, it093CSITF_2)
'... you just need **[of] some sports jerseys**, because the team...'

With these expressions we can, then, build the queries for extracting our target structures.

The basic structures of a noun phrase in European Portuguese, described in (2), can, finally, by captured by the following CQL expression, in (10).

(10) Noun phrase structure in CQL

```
([pos="P.*" & form=nform & form=fform & form=lex & form=reg & !contr])|( (([form="tod(o|a)s?" & pos="BQ.*" & form=nform & form=fform & form=lex]? [pos="DA.*|BD.*" & !contr & form=nform & form=fform & form=lex] [pos="BP.*" & form=nform & form=fform & form=lex]?) |([form!="tod(o|a)s?" & pos="BQ.*" & !contr & form=nform & form=fform & form=lex]))?([pos="R.*" & form=nform & form=fform & form=lex]{0,2} [pos="A.*" & form=nform & form=fform & form=lex])? [pos="E|N.*" & form=nform & form=fform & form=lex]([pos="R.*" & form=nform & form=fform & form=lex]{0,2} [pos="A.*" & form=nform & form=fform & form=lex])? ([pos="S.*" & form="d.*" & form=nform & form=fform & form=lex] (([form="tod(o|a)s?" & pos="BQ.*" & form=nform & form=fform & form=lex]? [pos="DA.*|BD.*" & form=nform & form=fform & form=lex] [pos="BP.*" & form=nform & form=fform & form=lex]?) | ([form!="tod(o|a)s?" & pos="BQ.*" & !contr & form=nform & form=fform & form=lex]))? ([pos="R.*" & form=nform & form=fform & form=lex]{0,2} [pos="A.*" & form=nform & form=fform & form=lex])? [pos="E|N.*" & form=nform & form=fform & form=lex] ([pos="R.*" & form=nform & form=fform & form=lex]{0,2} [pos="A.*" & form=nform & form=fform & form=lex])?)?)
```

These expressions can be quite long, and noun phrases can have as core elements pronouns and nouns, with different features concerning agreement, namely in what concerns Person features. Also, different types of nouns, common and proper nouns, occur in different structures. For these reasons, we decided to divide the queries into four separate expressions accounting for:

i) subject-verb agreement with personal pronoun,

ii) subject-verb agreement with common noun,

iii) subject-verb agreement with proper noun, and

iv) subject-verb agreement with relative pronoun.

Table 5, below, presents the CQL expressions for each of them.

Table 5. CQL expressions for target structures concerning subject-verb agreement (results in September 2023)

Nr.	Target structure / Proficiency level	CQL expression	Results
19	Subject-verb agreement -personal pronoun /B2	([pos="P.*S1" & form=nform & form=fform & form=lex & form=reg & !contr] [pos="R.*" & form=nform & form=fform & form=lex]{0,2} [pos="K.*" & form=nform & form=fform & form=lex & form=reg]? [pos="V.*1S" & lemma!="haver" & form=nform & form=fform & form=lex & form=reg]|[pos="P.*S2" & form=nform & form=fform & form=lex & form=reg & !contr] [pos="R.*" & form=nform & form=fform & form=lex]{0,2} [pos="K.*" & form=nform & form=fform & form=lex & form=reg]? [pos="V.*2S" & lemma!="haver" & form=nform & form=fform & form=lex & form=reg]|[pos="P.*S3|PI.*S" & form=nform & form=fform & form=lex & form=reg & !contr] [pos="R.*" & form=nform & form =fform & form=lex]{0,2} [pos="V.*3S" & lemma!="haver" & form=nform & form=fform & form=lex & form=reg]|[pos="P.*P1" & form=nform & form=fform & form=lex & form=reg & !contr] [pos="R.*" & form=nform & form=fform & form=lex]{0,2} [pos="K.*" & form=nform & form=fform & form=lex & form=reg]? [pos="V.*1P" & lemma!="haver" & form=nform & form=fform & form=lex & form=reg]|[pos="P.*P2" & form=nform & form=fform & form=lex & form=reg & !contr] [pos="R.*" & form=nform & form=fform & form=lex]{0,2} [pos="K.*" & form=nform & form=fform & form=lex & form=reg]? [pos="V.*2P" lemma!="haver" & form=nform & form=fform & form=lex & form=reg]|[pos="P.*P3|PI.*P " & form=nform & form=fform & form=lex & form=reg & !contr] [pos="R.*" & form=nform & form=fform & form=lex]{0,2} [pos="K.*" & form=nform & form=fform & form=lex & form=reg]? [pos="V.*3P" lemma!="haver" & form=nform & form=fform & form=lex & form=reg]) :: match.text_proficiency = "B2"	226
20	Subject-verb agreement common noun /B2	(([pos="N.S*" & form=nform & form=fform & form=lex]([pos="R.*" & form=nform & form=fform & form=lex]{0,2} [pos="A.S*" & form=nform & form=fform & form=lex])? ([pos="S.*" & form="d.*" & form=nform & form=fform & form=lex] (([form="tod(o|a)s?" & pos="BQ.*" & form=nform & form=fform & form=lex]? [pos="DA.*|BD.*" & form=nform & form=fform & form=lex] [pos="BP.*" & form=nform & form=fform & form=lex]?) | ([form!="tod(o|a)s?" & pos="BQ.*" & !contr & form=nform & form=fform & form=lex]))? ([pos="R.*" & form=nform & form=fform & form=lex]{0,2} [pos="A.*" & form=nform &	1228

		form=fform & form=lex])? [pos="E|N.*" & form=nform & form=fform & form=lex] ([pos="R.*" & form=nform & form=fform & form=lex]{0,2} [pos="A.*" & form=nform & form=fform & form=lex])?)? [pos="R.*" & form=nform & form=fform & form=lex]{0,2} [pos="K.*" & form=nform & form=fform & form=lex & form=reg]? [pos="V.*3S" & lemma!="haver" & form=nform & form=fform & form=lex & form=reg]) | ([pos="N.P*" & form=nform & form=fform & form=lex]([pos="R.*" & form=nform & form=fform & form=lex]{0,2} [pos="A.P*" & form=nform & form=fform & form=lex])? ([pos="S.*" & form="d.*" & form=nform & form=fform & form=lex] (([form="tod(o|a)s?" & pos="BQ.*" & form=nform & form=fform & form=lex]? [pos="DA.*|BD.*" & form=nform & form=fform & form=lex] [pos="BP.*" & form=nform & form=fform & form=lex]?) | ([form!="tod(o|a)s?" & pos="BQ.*" & !contr & form=nform & form=fform & form=lex]))? ([pos="R.*" & form=nform & form=fform & form=lex]{0,2} [pos="A.*" & form=nform & form=fform & form=lex])? [pos="E|N.*" & form=nform & form=fform & form=lex] ([pos="R.*" & form=nform & form=fform & form=lex]{0,2} [pos="A.*" & form=nform & form=fform & form=lex])?)? [pos="R.*" & form=nform & form=fform & form=lex]{0,2} [pos="K.*" & form=nform & form=fform & form=lex & form=reg]? [pos="V.*3P" & lemma!="haver" & form=nform & form=fform & form=lex & form=reg])):: match.text_proficiency = "B2"	
21	Subject-verb agreement proper noun /B2	([pos="TMS" & form=nform & form=fform & form=lex]? [pos="E" & form=nform & form=fform & form=lex]{1,3} ([form="de|do|da|dos|das" & form=nform & form=fform & form=lex] [pos="TMS" & form=nform & form=fform & form=lex]? [pos="E" & form=nform & form=fform & form=lex]{0,3})? [pos="R.*" & form=nform & form=fform & form=lex]{0,2} [pos="K.*" & form=nform & form=fform & form=lex & form=reg]? [pos = "V.*3S" & lemma!="haver" & form=fform & form=lex & form=reg]) :: match.text_proficiency = "B2"	206
22	Subject-verb agreement relative pronoun /B2	(([pos="N.S*" & form=nform & form=fform & form=lex]([pos="R.*" & form=nform & form=fform & form=lex]{0,2} [pos="A.S*" & form=nform & form=fform & form=lex])? ([pos="S.*" & form="d.*" & form=nform & form=fform & form=lex] (([form="tod(o|a)s?" & pos="BQ.*" & form=nform & form=fform & form=lex]? [pos="DA.*|BD.*" & form=nform & form=fform & form=lex] [pos="BP.*" & form=nform & form=fform & form=lex]?) | ([form!="tod(o|a)s?" & pos="BQ.*" & !contr & form=nform & form=fform & form=lex]))? ([pos="R.*" & form=nform & form=fform &	255

```
form=lex]{0,2} [pos="A.*" & form=nform &
form=fform & form=lex])? [pos="E|N.*" &
form=nform & form=fform & form=lex] ([pos="R.*" &
form=nform & form=fform & form=lex]{0,2}
[pos="A.*" & form=nform & form=fform &
form=lex])?)? [form="que" & form=nform &
form=fform &form=lex] [pos="R.*" & form=nform &
form=fform & form=lex]{0,2} [pos="K.*" &
form=nform & form=fform & form=lex & form=reg]?
[pos="V.*3S" & lemma!="haver" & form=nform &
form=fform & form=lex & form=reg]))|
(([pos="N.P*" & form=nform & form=fform &
form=lex]([pos="R.*" & form=nform & form=fform &
form=lex]{0,2} [pos="A.P*" & form=nform &
form=fform & form=lex])? ([pos="S.*" & form="d.*"
& form=nform & form=fform & form=lex]
(([form="tod(o|a)s?" & pos="BQ.*" & form=nform &
form=fform & form=lex]? [pos="DA.*|BD.*" &
form=nform & form=fform & form=lex] [pos="BP.*" &
form=nform & form=fform & form=lex]?) |
([form!="tod(o|a)s?" & pos="BQ.*" & !contr &
form=nform & form=fform & form=lex]))?
([pos="R.*" & form=nform & form=fform &
form=lex]{0,2} [pos="A.*" & form=nform &
form=fform & form=lex])? [pos="E|N.*" &
form=nform & form=fform & form=lex] ([pos="R.*" &
form=nform & form=fform & form=lex]{0,2}
[pos="A.*" & form=nform & form=fform &
form=lex])?)? [form="que" & form=nform &
form=fform &form=lex] [pos="R.*" & form=nform &
form=fform & form=lex]{0,2} [pos="K.*" &
form=nform & form=fform & form=lex & form=reg]?
[pos="V.*3P" & lemma!="haver" & form=nform &
form=fform & form=lex & form=reg])):
match.text_proficiency = "B2"
```

The expressions presented here allow us to extract subject-verb agreement cases such as the ones presented (11), ranging from simple sequences of pronoun-verb to long and more complex expressions involving MWE proper nouns with treatment forms, complex noun phrases, and relative pronouns with some distance to the core noun of the noun phrase they refer to.

(11a) ... mas **ele continuou** e finalmente chegou ao destino... (COPLE2, zh059CAATI_1)
'... but **he continued** and finally reached the destination...'

(11b) Quando **alguém parece** que precisa alguma coisa... (COPLE2, ja023CAATF)
'When **someone seems** to need something...'

(11c) ... recomendo que o **Governo de Portugal procure** maneiras adicionais... (COPLE2, en091CVATF)
'... I recommend that the **Government of Portugal looks** for additional ways...'

(11d) ... mas a **superficialidade da nova comunicação rápida não facilita** conversas sobre temas importantes... (COPLE2, de007CVATD)
'... but the **superficiality of the new rapid communication does not facilitate** conversations on important topics...'

(11e) Acredito que muitas **caraterísticas dos portugueses hoje são** relacionadas com a História Portuguesa. (COPLE2, zh065CAATF)
'I believe that many **characteristics of the Portuguese today are** related to Portuguese History.'

(11f) Os **exemplos dessa injustiça encontram-se** por toda a parte, ... (COPLE2, ru013CAATF_3)
'The **examples of this injustice are** everywhere, ...'

(11f) ... são as pequenas **coisas do dia-a-dia que fazem** a diferença e o enriquecimento cultural... (COPLE2, de016CVATI)
'... it's the **little everyday things that make** a difference and cultural enrichment... '

(11g) A solidão, o stresse e a depressão são as **doenças das sociedades modernas que causarão** mais prejuízos, humanos e económicos, no seculo XXI. (COPLE2, zh007CVATF)
'Loneliness, stress and depression are the **diseases of modern societies that will cause** the most human and economic damage in the 21st century.'

(11h) ... temos o exemplo dos **governos de extrema esquerda que se transformaram** em ditaduras nos países do Leste da Europa. (COPLE2, it015CVATF)
'... we have the example of **far-left governments that turned** into dictatorships in Eastern European countries.'

Before discussing, in more detail, the analysis and evaluation of the results achieved, it is necessary to refer some practical decisions taken. The first one concerns the option to not consider the material that can occur before the core noun in the subject noun phrase. We decided not to consider this material since it would make the expressions lengthier and, with that, with more possibility of errors. Additionally, it is the core noun that determines the agreement features. Also, the internal structure of the proper noun phrase is different from the one determined for noun phrases in general. This option is related to the fact that many proper nouns are tagged with the E tag in the corpus, instead of other part-of-speech tags (e.g., *Banco/E Alimentar/E* and not *Banco/N Alimentar/A*). So, the CQL expression considers sequences of adjacent proper nouns, mediated or not by the preposition *de*, instead of searching for regular sequences expressing regular noun phrase structures. Finally, all the expressions, as presented in (3), consider the possibility of adverbs and clitics occurring between the subject and the verb. The examples in (12) illustrate these cases, highlighted in italics.

(12a) Por exemplo, ele *não* sabia que... (COPLE2, en014CAATD)
'For example, he did *not* know that...'

(12b) Eu *pessoalmente* vejo e acredito que... (COPLE2, zh018CVATF)
' I *personally* see and believe that...'

(12c) Contudo, é sempre importante para os consumidores *não se* deixarem... (COPLE2, nl010CVATI)
'However, it is always important for consumers *not to* [*themselves*] let...'

(12d) Se vocês *não me* pagarem até Abril, vou... (COPLE2, ja019CAATD)
'If you do *not* [*me*] pay by April, I'll...'

3. Analysis and evaluation of results

To evaluate the results, we conducted a manual revision of the cases extracted. Table 6 presents the evaluation of the results in terms of number of extracted occurrences, number of bad results, i.e., occurrences that do not correspond to the target structure queried, and the percentage of good results obtained.

Table 6. Evaluation of results (September 2023)

Target structure / Proficiency level		Query Nr.	Number of extracted occurrences	Bad results	Percentage of good results
<ç> with [s] value / A2		1	345	0	100%
<x> with [z] value / B2		2	145	0	100%
Stress marking / C1		3	2 940	0*	100%
Nasality / A2		4	3 596	0*	100%
Simple Conditional/C1		5	55	0	100%
Future with Auxiliary *ir* / A2		10	191	1	99,5%
Subjunctive *Pretérito Perfeito Composto* / C1		11	2	0	100%
Connectors	Atomic / A1	6	115	0	100%
	Atomic / B1	7	7	0	100%
	MWE / B1	12	33	1	97%
	MWE / B1	13	5	1	80%
Academic life vocabulary	Atomic / A1	8	485	0	100%
	MWE /A1	14	0	0	-
	MWE / A1	15	9	9	100%
Urban space vocabulary	Atomic / B2	9	299	12	96%
	MWE /B2	16	3	3	100%
	MWE / B2	17	0	0	-
	MWE / B2	18	0	0	-
Subject-verb agreement/B2	personal pronoun	19	226	0	100%
	common noun	20	1 228	6*	98%
	proper noun	21	206	25	91,9%
	relative pronoun	22	255	4	98,4%

*The evaluation was performed over a sample of 25% of cases randomly selected.

Our overall evaluation is that the CQL expressions built allow the extraction of very reliable results. In 22 CQL expressions, we only performed below 98% in four cases, of these only one below 91,9%. Besides assessing the precision of the results, we also analysed the bad results to understand possible reasons for their occurrence and how to address these, if possible.

Several reasons account for most of the bad results, as illustrated below.

i. Errors in the part-of-speech tag.

(13) ... **visto**_CommonNoun **que é** imenso facil para saber como foi passado... (COPLE2, zh022CVA1TF)
'... **since**_CommonNoun it is really easy to find out how it was spent...
(bad result from CQL expression 22: Subject-verb agreement/B2 relative pronoun)

ii. Missing annotation.

(14a) ... **aquelos**_missing normalized form **que mais tiverom**_missing normalized form suceso. (COPLE2, it101CSATF)
'... **thoso**_ missing normalized form that **more hod**_missing normalized form succes.

(bad result from CQL expression 22: Subject-verb agreement relative pronoun/B2)

(14b) ... **apesar de ninguém pode**_missing normalized form ser o "superman"... (COPLE2, zh044CAMTI)
'... **although no one can**[inflected form]_missing normalized form be "superman"...'
(bad result from CQL expression 13: Complex connectors/B1)

iii. Linear order coincidences related to the annotation of locutions.

(15a) De **maneira**_CommonNounSingular **que**_RelativePronoun **faça**_VerbSingular isto, temos dois passos... (COPLE2, zh013CVATD)
'In such a **manner**_CommonNounSingular **that**_RelativePronoun [it] **does**_VerbSingular this, we have two steps...'
(bad result from CQL expression 22: Subject-verb agreement relative pronoun/B2)

iii. Linear order coincidences related to the recursive nature of noun phrase and prepositional phrase structures.

(15b) ... por isso cualquer problema_CoreNoun de uma **pessoa do povo é** considerado... (COPLE2, es036CVATF)
'... for that any problem_CoreNoun of a **person of the people is** considered...'
(bad result from CQL expression 20: Subject-verb agreement common noun/B2)

iv. Linear order coincidences related to sentence subjects.

(16a) ... [ter o tempo para ganhar mais **dinheiro**]Subject **é** mais importante. (COPLE2, zh007CVATF)
'... [having the time to earn more **money**]Subject **is** more important.'
(bad result from CQL expression 20: Subject-verb agreement common noun/B2)

(16b) ... [viver no **estrangeiro**]Subject **é** como uma faca de dois grumes... (COPLE2, zh081CVATI)
'... [living in a foreign **country**]Subject **is** like a double-edged knife...'
(bad result from CQL expression 20: Subject-verb agreement common noun/B2)

iv. Linear order coincidences related to omitted subjects.

(17a) ... e "roubaram" todas as **riquezas que** [Omitted subject 'eles'] **lá tinham**. (COPLE2, es032CVATF)
'... and "stole" all the riches that '[Omitted subject 'they'] **there had**.
(bad result from CQL expression 22: Subject-verb agreement relative pronoun/B2)

(17b) ... a **língua que** [Omitted subject 'ele'] **fala**, a tradição que segui,... (COPLE2, zh077CVATF)
'... the **language that** [Omitted subject 'he'] **speaks**, the tradition [I] followed, ...
(bad result from CQL expression 22: Subject-verb agreement relative pronoun/B2)

vi. Linear order coincidences related to indefinite subject constructions with *-se*

(18a) ... o mesmo, com a **diferença que hoje se**_clitic **migra** para viver, porque... (COPLE2, es087CVATI1)
'... the same, with the **difference that today one migrates** to live, because...'
(bad result from CQL expression 22: Subject-verb agreement relative pronoun/B2)

(18b) ... , com a **esperança que não se**_clitic **volte** atrás e que não sejam... (COPLE2, it101CSATF)
'..., with the **hope that one does not go** back and that [they] do not be...'
(bad result from CQL expression 22: Subject-verb agreement relative pronoun/B2)

vii. Lexical ambiguity

(19a) Nós gostamos de ser o **centro** do mundo, ... (COPLE2, zh017CVATF)
'We like to be the **center** of the world...'
(bad result from CQL expression 9: Urban space vocabulary/B2)

(19b) ... gente que vive na **rua** e que não tem... (COPLE2, de026CVATD)
... people that live in the **street** and that does not have...'
(bad result from CQL expression 9: Urban space vocabulary/B2)

The bad results listed here can be expected considering the fact that i) we are dealing with automatically tagged data, which means that even with high scores, there is always a small margin for error; ii) we are dealing with data annotated only at the part-of-speech level, with no syntactic analysis or parsing; iii) and we are dealing with natural language data that carries structural and lexical ambiguity. For these reasons, and considering the percentage of good results achieved, we feel quite confident that the CQL expressions here devised are efficient, robust, and sufficiently motivated.

4. Final remarks

The CQL expressions proposed in this paper allow for the extraction of the target structures aimed at with very good results. Besides being ready to use, these expressions can also be easily adapted to accommodate different sets of vocabulary (e.g., in CQL expressions 8 and 9, for instance), or to select for different levels of proficiency depending on the curricula in use (e.g., changing the value of the 'match.text_proficiency' attribute).

Although not representing an innovative perspective on the data or on the process - CQL and regular expressions have been widely used since many decades -, it is our belief that having linguistically motivated expressions, with high degree of precision, can be very useful for researchers working in LE and FL acquisition, in phenomena of linguistic interference, and analysis and diagnosis of LE/L2 proficiency levels. For instance, besides extracting data by specific proficiency levels, it is possible to correlate level of proficiency, mother tongue of the learner and the occurrence of target structures and/or errors. Moreover, the work here depicted provides usable query expressions that are not easily built using the COPLE2 Query Builder. As shown in Figure 3, the options available to the user do not cover the range of attributes and of operators used in the CQL expressions we propose here.

The description of the process presented here also contributes to further enhancements, either in what concerns the development of other CQL expressions to extract other phenomena, either in what concerns the development of the COPLE2 query interface. For instance, the queries presented here can be added to the COPLE2 web interface as pre-built queries. This way, the users could use these queries immediately and would not have to build the CQL expressions from scratch.

The possibility of extracting data not explicitly marked in the corpus, such as the target structures considered here, potentiates the use of COPLE2 corpus for studying these types of structures in terms of frequency of occurrence, frequency by proficiency level, relation between target structures and mother tongue (for instance, adding the attribute `:: match.text_mothertongue` to the CQL expressions to further restrict the queries), as well as testing and validating hypothesis in larger data sets. It can also easily and immediately contribute to the building of didactic materials based on real data and real difficulties.

Acknowledgments

We are very grateful to the anonymous reviewers for their comments and suggestions.
We thank Joana Oliveira for her involvement and valuable input in the earlier stages of this work.

Funding

Part of this research is supported by the Portuguese national funding through the FCT – Portuguese Foundation for Science and Technology, I.P. as part of the project UIDB/LIN/03213/2020 and UIDP/LIN/03213/2020 – Linguistics Research Centre of NOVA University Lisbon (CLUNL).

References

Amaro, Raquel, Susana Correia, Carolina Gramacho & Amália Mendes (2020) Automatização no diagnóstico de nível de língua: Anotação e versatilidade dos recursos. *Revista da Associação Portuguesa de Linguística* (7), pp. 1–20. https://doi.org/10.26334/2183-9077/rapln7ano2020a1

Cambridge (2012) Cambridge Sketch Engine - Using Corpus Query Language (CQL) (1.3). Cambridge University Press. Available at https://www.cambridge.org/sketch/help/userguides/CQL%20Help%201.3.pdf

Castello, Erik, Katherine Ackerley & Francesca Coccetta (eds.) (2016) *Studies in learner corpus linguistics*. Peter Lang.

Christ, Oli (1994) A modular and flexible architecture for an integrated corpus query system. In *Proceedings of COMPLEX'94: 3rd Conference on Computational Lexicography and Text Research*, pp. 23–32. https://doi.org/10.48550/arXiv.cmp-lg/9408005

del Río, Iria, & Amália Mendes (2018) Error annotation in a Learner Corpus of Portuguese. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation* (LREC 2018). European Language Resources Association (ELRA), pp. 4116–4119. Available at https://aclanthology.org/L18-1

Gramacho, Carolina, Ana Madeira, Cláudia Martins, Nélia Alexandre, Jorge Pinto & Susana Correia (2019) Por nível: Construção e validação de um teste de colocação para o Português Língua Estrangeira–resultados de um estudo-piloto. *Revista da Associação Portuguesa de Linguística* (5), pp. 172–189. https://doi.org/10.26334/2183-9077/rapln5ano2019a13

Mendes, Amália, Sandra Antunes, Marteen Janssen & Anabela Gonçalves (2016) The COPLE2 Corpus: A learner corpus for Portuguese. In *Proceedings of the 10th Language Resources and Evaluation Conference – LREC'16*, pp. 3207–3214. Available at https://aclanthology.org/L16-1

Talhadas, Rui (2016) Mapping grammatical structures onto proficiency levels. In *Proceedings of 12th International Conference on Computational Processing of the Portuguese Language (PROPOR 2016)*. Available at http://propor2016.di.fc.ul.pt/wp-content/uploads/2016/07/RuiTalhadasPROPORSRW2016.pdf

Tracy-Ventura, Nicole & Magali Paquot (eds.) (2020) The Routledge handbook of second language acquisition and corpora. Routledge.