HOW TO FIND A SHINING NEEDLE IN THE HAYSTACK. QUERYING COROLA: SOLUTIONS AND PERSPECTIVES

DAN CRISTEA[1], NILS DIEWALD[2], GABRIELA HAJA[3], CĂTĂLINA MĂRĂNDUC[4], VERGINICA BARBU MITITELU[5], MIHAELA ONOFREI[6]

Abstract. The present paper examines a variety of ways in which the Corpus of Contemporary Romanian Language (CoRoLa) can be used. A multitude of examples intends to highlight a wide range of interrogation possibilities that CoRoLa opens for different types of users. The querying of CoRoLa displayed here is supported by the KorAP frontend, through the querying language Poliqarp. Interrogations address annotation layers, such as the lexical, morphological and, in the near future, the syntactical layer, as well as the metadata. Other issues discussed are how to build a virtual corpus, how to deal with errors, how to find expressions and how to identify expressions.

Keywords: CoRoLa, Poliqarp, KorAP, corpus querying, lexical level, morphological level, syntactical level, metadata, virtual corpus.

1. INTRODUCTION

In this paper, we deal with the use of CoRoLa, showing how it can be queried from different perspectives by potential users, driven in their searches by a diversity of motivations. The corpus includes, for the time being, only texts more recent than 1945, therefore it should be considered a contemporary corpus. Motivations for using CoRoLa go beyond a lexicographic interest, although from a lexicographic perspective, updating dictionaries is a continuous endeavour, as a language is continuously mutating, evolving, aging, etc. CoRoLa can be of use in learning Romanian as a foreign language, when teaching Romanian in schools, when looking at uses and searching for usage errors, when examining contemporary exaggerations of technical jargon.

[1] "Alexandru Ioan Cuza" University of Iaşi, Faculty of Computer Science; Iaşi branch of the Romanian Academy, Institute for Computer Science, dcristea@info.uaic.ro

[2] Leibniz-Institut für Deutsche Sprache, Mannheim, diewald@ids-mannheim.de

[3] Iaşi Branch of the Romanian Academy, "A. Philippide" Institute for Romanian Philology, gabihaja@gmail.com

[4] Romanian Academy, "Iorgu Iordan – Al. Rosetti" Institute of Linguistics; "Alexandru Ioan Cuza" University of Iaşi, Faculty of Computer Science, catalinamaranduc@gmail.com

[5] Romanian Academy, Research Institute for Artificial Intelligence, Bucharest, vergi@racai.ro

[6] Iaşi Branch of the Romanian Academy, Institute for Computer Science, mihaela.plamada. onofrei@gmail.com

RRL, **LXIV**, *3*, p. 279–292, Bucureşti, 2019

This paper is not intended as a user manual for CoRoLa's frontend, nor for any of its query languages. By selecting a sum of examples, which we considered interesting, by grouping them in categories and sometimes ranking them from simple to more complex, we wanted to inspire and attract to CoRoLa different categories of potential users, with no intention of describing all the possible ways of querying it.

2. POLIQARP

KorAP (Bański *et al.* 2013), one of the search frontends of CoRoLa, displays a number of query languages for the interested user: Poliqarp, Cosmas II, Annis QL, CQL v1.2 and FCSQL. The examples we put in evidence in this paper only use Poliqarp (Przepiórkowski *et al.* 2004), a query language for searching large linguistic data sets[7]. This language is a variant of CQP (Christ 1994), and it is the most popular of the query languages supported by KorAP. Its popularity is due to a combination of attractive features[8], among which the fact that it can handle not only raw texts but also tagged corpora, in which tagging can take any form (CoRoLa annotated texts and their metadata, as described in Tufiș *et al.* 2019, in this volume, are encoded in XML). Poliqarp is based on *regular expressions*[9], thus allowing one to formulate from very simple to very sophisticated conditions involving sequences of words. Conditions can exploit the internal structure of the tags, they do not depend on a particular tagset and can be used on corpora of texts written in the native scripts of almost any language, if it is encoded in the UTF-8 format. Implementations based on Poliqarp are generally quick in providing an output but depending on the size of the corpus and on the complexity of the query, the waiting time may vary from several seconds to a minute. The KorAP query interface (Diewald and Margaretha 2017) implements an extension of Poliqarp, named Poliqarp+.

3. QUERYING

3.1. The lexical level

In this section, we will show the first and simplest level of queries: looking for word occurrences. In Poliqarp, the elementary units of a query are called segments, and, in most cases, these refer to words. The result of a search[10] is case sensitive to typed letters, so

[7] By Instytut Podstaw Informatyki Polskiej Akademii Nauk (IPI PAN; Institute of Computer Science, Polish Academy of Sciences, www.ipipan.waw.pl). Poliqarp is a free/open source software, available under the terms of the GNU General Public License.

[8] http://poliqarp.sourceforge.net/about.html

[9] A comprehensive tutorial on regular expressions can be found at https://www.datacamp.com/community/tutorials/python-regular-expression-tutorial. For space constraints, in this paper we give only short comments on the regular expressions occurring in our examples, to make them easily understandable by anyone.

[10] For lack of space, we did not include in this paper print screens showing the results of queries, but the interested user could simply reproduce the expressions we give in the CoRoLa frontend and visualize the outputs.

inputting copac (‘tree’), or Copac, or COPAC will output different contexts. To elude case sensitivity, the word searched for should be followed by /i, as here in: copac/i. Similar results can be obtained by using a notation in which segments in queries are delimited by *square brackets*. These are called *complex segments*. To query CoRoLa, annotations in bracketed segments have to be prefixed by drukola/, thus naming the so-called *foundry*.

In a bracketed (complex) segment, additional constraints can be added on the term under scrutiny by providing *key*=*value* pairs. For instance, the key orth goes for surface forms. Other keys supported by Poliqarp and CoRoLa are base (for lemmas) and pos (for parts of speech). Because KorAP is limited, for the time being, to a single tokenization, the foundry prefix can be omitted for orth. As such, [orth=copac] brings the same results as with the simple query copac, and [orth=copac/i] – with the simple query copac/i. The expression [drukola/base=copil] searches for words with lemma *copil* (‘child’), while [drukola/pos=verb] brings all occurrences of verbs from the corpus. As both lemma and part of speech can be annotated in multiple foundries, defining the prefix in queries is recommended, and, when missing, it will be automatically replaced by a default foundry, as configured by the system.

Regular expressions (REs) in segments shall be placed in-between double quotes. Regular expressions can directly refer to forms of words or any other key values. Following are some examples of uses of REs:

- "hip"/x, same as [drukola/orth="hip"/x] - looks for words that contain the sequence of letters *hip* anywhere in the word;
- "hiper.+", same as [drukola/orth=”hiper.+”], looks for occurrences of words starting with *hiper*; this is because the . (dot) sign matches any symbol and the + (plus) sign is the Kleene operator obliging for at least one occurrence;
- "hiper.+"/i or [drukola/orth="hiper.+"/i] - same as above, but case insensitive;
- ".+oai.+" or [drukola/orth=".+oai.+"] - words that contain the sequence *oai* somewhere in the middle;
- ".+oai.+"/i or [drukola/orth=".+oai.+"/i] - same as above, but case insensitive;
- ".+tor" or [drukola/orth=".+tor"] - words ending with the sequence -*tor*.

Within segments, logical conditions can be formulated. For instance, [drukola/orth="copi.+" & drukola/pos=verb] will output verbs starting with *copi-*.

Similarly, any of the queries [drukola/orth=acele & drukola/pos=noun] or [drukola/orth=acele & drukola/base=ac] search for the form *ácele* (‘needles’), i.e. plural forms of the noun *ac* (‘needle’). Forms accented *acéle* (‘those’), i.e. as a demonstrative determiner, can be obtained with the query [drukola/orth=acele & drukola/pos=determiner].

Two (or more) words in immediate text vicinity can be obtained by placing them in sequence in the query expression, for instance: copil cuminte, or [drukola/orth=copil] [drukola/orth=cuminte][11].

To express the negation of a value, the sign “!” (exclamation mark) should prefix the “=” (equal) sign, making thus a *key*!=*value* pair. For instance, [drukola/base=putea] [drukola/base!=să] will retrieve all occurrences in which the modal *putea* (‘can, may’, infinitive) is not followed immediately by the particle *să* (subjunctive particle, unmarked for voice in this context).

[11] Since in the present implementation punctuation is ignored, in the displayed output punctuation marks could appear interposed between the two words.

Searching for co-occurring words placed at a distance can be done by using the *empty segment* symbol: []; it skips any word (not also punctuation). For instance: the expression [drukola/base=copil] [] [drukola/base=cuminte] brings contexts in which *copil* ('child') and *cuminte* ('good, quiet, unspoiled') are placed at one-word distance. Curly brackets can pair the empty segment to indicate different ranges of skipped elements in the retrieved contexts, for instance: [drukola/base=copil] []{2} [drukola/base=cuminte] - *copil* and *cuminte* are placed at exactly 2 words distance, ... []{2, } ... - for at least two interposed words; ... []{2,5} ... - for a minimum of 2 and a maximum of 5 interposed words.

3.2. The morphological level

Consider the case of the homograph *vesélă* (noun indef. sg. dir.[12]) - *véselă* (adj. fem. sg. dir. indef.) ('dishes, tableware' vs 'merry'). When searching for this form in the corpus, we get contexts for both the noun and the adjective. In such cases, the user needs to add some morphological restrictions to the query so that the contexts of the word of interest are displayed.

For the beginner user, the most accessible way of doing this is by finding an example of interest in the corpus by means of the lexical query, then expand the Tokens view for it, and click on the morphological information that helps adjust the query. This will automatically create a box starting with New Query:, immediately followed by the selected morphological information. Clicking on another element in the annotation field will add it to the New Query box[13]. When the user considers that all restrictions have been added to this field, (s)he can submit the query to the main query box (the one on top of the page) by a mere click anywhere in the box New Query. Once the query phrase is here, it is necessary to either click on the search button or position the cursor in this main query box and press Enter. The query is sent to the system and the results are displayed. In Figure 1 we exemplify this functioning for the contexts of the noun *veselă* in the singular direct case. We can see that one of the concordances is expanded (by a mere click on it), the Tokens view is also expanded (also by clicking it), the lemma *veselă* (see layer l), the case:direct and number:singular were added to the new query (also by clicking them), because they are the necessary and sufficient elements that ensure the retrieval of only the contexts of the noun form. As we did not specify the definiteness of the noun, the user should expect both definite (*vesela*) and indefinite (*veselă*) forms as results. If only the indefinite one is required, the query can be further refined by adding to it the definiteness:no restriction.

Examples can also be used to create new queries, not only to refine ones. A query can also be created starting from a previous concordance and grouping elements from all layers: inflected forms (from the top layer), lemmas (from layer l), morphological information (from layer m), parts of speech (from layer p).

All morphological tags that are used for annotating CoRoLa are described on the project website, www.corola.racai.ro[14]. In the section Interogare, links to three (downloadable) files are available: *Specificații privind etichetele MSD* (Specifications on

[12] Romanian has a merged Nominative-Accusative, a syncretic Genitive-Dative and a Vocative case system. Nom-Acc is labeled in the Specifications as direct case, Gen-Dat as oblique case.

[13] A second click on the same element will automatically delete it from the New Query box.

[14] See also Tufiş *et al.* 2019, in this volume.

the morphosyntactic description tags), *Specificații privind etichetele CTAG* (Specifications on the category tags), *Specificații privind codificarea de tip trăsătură-valoare utilizată de KorAP* (Specifications on the attribute-value encoding used by KorAP). The first file contains the description, in English and Romanian, and examples of all the tags that appear in the Tokens view after the attribute msd. The second file contains the description, in English and Romanian, of the values of the attribute ctag in the Token view. The third one contains the equivalence between the MSD tags and the attribute-value format. Consequently, there is a redundancy in the morphological layer: both the MSD tag and the attribute:value representation offers the users the same information; while the latter is friendly, the former is rather esoteric.

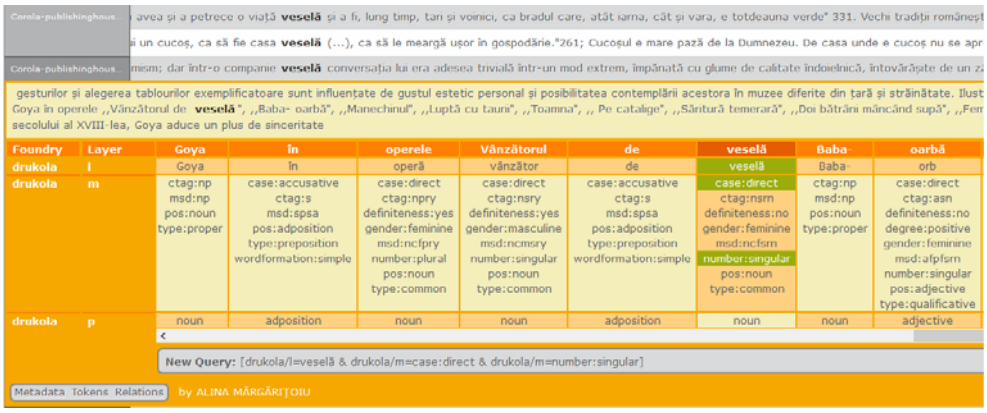

Fig. 1. Using morphological information from already found examples to refine queries.

All queries must start with the foundry in which the search will be made: [drukola/m=]. We exemplify now several types of searches. The following queries retrieve sequences of words of the type: indefinite singular common noun at a direct case + preposition + definite plural common noun at a direct case[15]. The first query is written with the help of MSDs, the second with the help of attribute-value pairs[16] and the last one with the help of CTAGs:

(1) [drukola/m="msd:nc.srn"][drukola/m="msd:s.*"] [drukola/m="msd:nc.pry"]

(2) [drukola/m=pos:noun & drukola/m=type:common & drukola/m=number:singular & drukola/m=case:direct & drukola/m=definiteness:no] [drukola/m=pos:adposition] [drukola/m=pos:noun & drukola/m=type:common & drukola/m=number:plural & drukola/m=case:direct & drukola/m=definiteness:yes]

(3) [drukola/m=ctag:nsrn][drukola/m=ctag:s] [drukola/m=ctag:npry]

[15] The unspecified gender is represented by the dot (any character in regular expression) in the MSD tag.

[16] This type of search is somehow slower.

3.3. The syntactic level

In this section we bring forward a sum of syntactic issues, before the implementation of this level for CoRoLa. The motivation for this look-ahead presentation is twofold: to inform the users of future possibilities of querying the syntactic level, but also to wave a flag in front of the eyes of our team members towards an implementation that, in the near future, would fully allow performing searches displaying the use of all these targeted features.

A popular representation in computational linguistics is provided by the linguistic dependency model, originally proposed by Tesnière (1959), which mainly postulates that a syntactic structure consists of binary asymmetric relations between lexical items. A relation connects a head (governor, regent) with a dependent (complement, adjunct, subordinate). These dependencies are represented by labelled graph structures. The overall structure of a sentence has the form of a tree, as each lexical item is restricted to have one single head (except for the root of the structure). Segmentation techniques, like those intended to recognise groups of lexical items (for instance, NPs, VPs, etc.), usually dealt with regular expressions, can be seen as cutting operations on dependency hierarchies. The notion of valency (inventory of obligatory dependents) of a lexical item is a concept used for describing cognitive processes in syntactic structures. By this, syntactic descriptions are naturally glued to semantic descriptions, and the richness of the range of relations and constraints configure specificities of a language. Recently, Universal Dependency (UD) has got a lot of attention from computational linguists by its attempt to simplify and unify grammars for a large range of languages (McDonald *et al.* 2013). If not manually annotated, a dependency structure is obtained by running a parser on the raw text, previously tagged for parts of speech. Elements of a syntactic query are lexical items, heads, phrases and/or syntactic relations.

As will be seen below, part of these queries aims at obtaining statistical data, triggered by specific search criteria. Statistical functions are part of the distributed framework of KorAP, which is still under development. For the time being, only statistics on corpus data and match counts can be obtained in the installed frontend, while other more specific types of statistics can only be collected by consulting the Web-API using external scripts[17].

a. What types of configurations of syntactic subordinates can a particular word have?

Questions of this type are important, for instance, when building dictionaries of verbal patterns. As such dictionaries (Levin 1993; Hanks 2013; Pană Dindelegan 1974; Barbu 2018) put in evidence typical syntactic and/or semantic structures for each verb[18], they are useful to both linguists and computational linguists. The patterns found could, for instance, constitute constraints on the dependencies attached to a particular word and incorporated into a parser.

b. What is the greatest number of subordinates that the verb *a cânta* can have?

Let's note that to properly answer this query, a future implementation should display a different type of result from mere occurrences, notably a summary of a statistical search, in which only the relations should be brought forward. Because this kind of result is

[17] For all features provided by the Web-API, consult the documentation at https://github.com/KorAP/Kustvakt/wiki.

[18] For English, see for instance: http://www.pdev.org.uk.

different from anything corpus searches can produce, it is envisioned that it could be part of a library of specific Web-APIs that receives the primary output of a corpus query and yields inventories of all kinds, among them the one requested. COSMAS 2, the predecessor of KorAP, already has implemented statistical result views, which will be integrated in KorAP as well. In the meantime, the workaround to gather all data and calculate the statistics in a separated tool is possible, using the Web-API via external scripts.

c. In general, it is accepted that Romanian has a relatively free word order, yet it is interesting to properly prove by examples the degree to which this is true and to signal cases when it is not free. This means, for instance, to detect a set of verbs for which a certain dependency appears obligatorily in front of / after the verb, and to see in which cases the change of word order can lead to semantic shifts.

An example where the word order changes the meaning of an adjective is *o* ***nouă*** *rochie* (‘another dress’) vs *o rochie* ***nouă*** (‘a new dress’) (for a dedicated study on this matter, see Cornilescu and Giurgea 2013). Instances of marked word order in Romanian occur, for instance, in noun – attributes, such as: ***al ei*** *gând*, ***al lui*** *copil* (‘her thought’, ‘his child’). The word order specific to Romanian, in this case, is noun + genitival attribute, with definite determiner function (*gândul ei*, *copilul lui*, ‘her thought’, ‘his child’), and no stylistic function. In the first 250 contexts of the form *al ei* (‘her(s)’) found in CoRoLa, all occurrences with the inverted, marked word order *al ei* +N appear in poetic contexts.

d. Are there limits for the distance a certain constituent can occur at with respect to the main verb?

These distances can be measured linearly, as offsets of words, in the surface strings, or by counting interposed subordinates. Psychologists, for example, could be interested to study correlations of this type, to measure the capacity of the human brain to store and systematize. Computational linguists could also include statistical data reported by these types of queries into their processing tools, in the form of confidence weights for attaching subordinates.

e. What is the range of head words that can be detected in the position of each syntactic relation of the verb *a cânta*?

Queries of this type can be triggered by interests to detect semantic patterns of verbs, for instance by generalising the range of detected words occurring on a certain verbal role to semantic classes. Also, significances encumbered by the word order can be triggered. Verbal syntactic subordinates are in close correlation with their semantic roles and an inventory of semantic roles of verbs has recently been created for Romanian language (Barbu 2018) in the spirit of similar resources for other languages, see (Levin 1993) for English or (Pala 1999) for Czech.

f. Examples of sentences, in which lemma *cânta* with msd = Vmp.* (participle) or msd = Vmg.* (gerund) has subordinates.

This query investigates the way in which the verbal mode influences the dependency pattern. Dependencies in participle and gerund are rare in language, however, they are required for exemplifications in various linguistic works. For example, in *Dicționarul limbii române* (DLR, *Dictionary of Romanian language*) the distinction between the participle and the homonymous adjective (obtained through conversion) is based on formal criteria.

Consequently, the adjective cannot have subordinates of the same type as the participle has, due to the adjective having lost its meaning as both action and the result of an action. An example of such can be *deschis, -ă* (‘*open’*, masc./fem. indef.), which, as an adjective, cannot have either direct or indirect object, circumstantial or agent complements; on the other hand, it has degrees of comparison. Compare the following CoRoLa examples:

– participle:

(1) Mama [...] trecea în cealaltă cameră, lăsând uşa **larg deschisă**
‘Mother [...] walked into the other room, leaving the door wide open’

– adjective:

(2) a. lista rămâne **deschisă** b. cu inima **deschisă**
‘the list remains open’ ‘with an open heart’

g. Sequences #1:pos=noun #2:pos=verb #3:pos=verb, such that #2 subordinates #3 and #3 subordinates #1 (here and following, by #1, #2, #3 we note syntactic constituents).

For instance, in the sentence *Elevii trebuie să frecventeze cursurile.* (‘Students have to attend the courses’), the head of *elevii* is *frecventeze*, whose head is *trebuie* (it would also be interesting to see if, in similar constructions, there exist occurrences of other verbs than *trebuie* in the position of root). Dependencies of this type, in which a descendent of a literal is placed in front of the head of that literal, can be called *advanced* and are of major importance in a dictionary of verbal patterns of the language.

3.4 Complex queries

What we call complex queries here are queries that use combinations of search levels. Some examples are commented below:

a. searching a word by lemma + morphology: [drukola/base=preşedinte & drukola/m=case:oblique & drukola/m=number:plural] will bring occurrences of the word *preşedinte* (‘president’) in genitive/dative plural,

b. searching two words by lemma and morphological features: [drukola/base=pian & drukola/m=case:direct & drukola/m=number:singular][drukola/m=case:oblique & drukola/m=type:proper] – looking for sequences of the word *pian* (‘piano’) in nominative/accusative plural followed by a proper name in genitive/dative,

c. searching in the same sentence for sequences of two words by morphological features: contains(<dereko/s=s>, [drukola/m=pos:adjective][drukola/m=type:proper & drukola/m=case:vocative & drukola/m=definiteness: yes]) – looking for sequences made of an adjective followed by a proper noun in vocative. The contains(<dereko/s=s>, ...) condition states that both tokens should be part of the same sentence.

4. MISCELLANEOUS

4.1 Building a virtual corpus

In addition to the queries illustrated in this paper, which basically search for words in context, KorAP also supports document queries, allowing to define a subcollection of documents to search in. This subcollection, also called a “virtual corpus” (Bański *et al.* 2013), can be defined by document-level metadata constraints, combined by logical operation. For implementation details, see Diewald and Margaretha (2017).

The simplest constraint is an equality relation with a metadata property, such as restricting the search to all documents with the textType field being BlogPost. Figure 2 shows an example of available metadata fields in CoRoLa, as displayed in the metadata view of KorAP (for more information on metadata in CoRoLa see Tufiş *et al.* 2019, in this volume). Relations with metadata properties can also be negated, defined as regular expression, or, depending on the field type, defined as being in a range of values, such as restricting the search to all documents with a pubDate (i.e. the publication date) after the year 2010. These constraints can be combined using logical and and or relations. KorAP provides a visual builder tool to simplify the creation of nested virtual corpora and choosing the right metadata fields (see Figure 3).

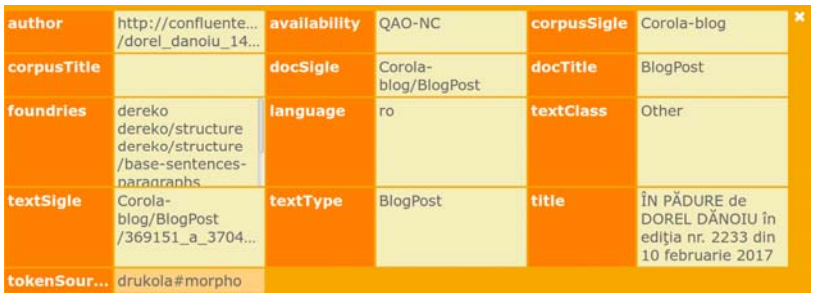

Fig. 2. Metadata view for a single document in the corpus. All fields with a light-yellow background can be used to represent a metadata constraint as part of a virtual corpus.

Fig. 3. The virtual corpus view in KorAP with statistical information about the corpus size.

Virtual Corpora can be stored and referenced by name as (nested) constraints in further virtual corpora. This makes it possible to refine stored virtual corpora in the interface without complex filtering. Stored and referenced virtual corpora also provide performance benefits, as the collection of matching documents are cached on the index level. This becomes significant when a virtual corpus consists of hundreds or thousands of metadata constraints, like the comparable corpus for the Romanian and the German language (cf. Kupietz *et al.* 2019, in this volume).

4.2 Dealing with errors

CoRoLa was built following a sophisticated processing protocol, as shown in Tufiş *et al.* (2019) and Gîfu *et al.* (2019), in this volume. All the component phases of the protocol were prone to errors (for instance, filling in metadata, cleaning the text, annotation chain, etc.). Many have been detected and corrected, mainly following some automatic procedures, but errors still exist. On the other hand, grammatical or usage errors in the corpus may be numerous as well, as in the selection of primary documents we did not apply any restrictions of language use. These are valuable for the study of language and should not be looked at from a prescriptive perspective.

In this section we will show a few examples of errors. Some of them are due to technology, more having to do with the ungrammatical use of language.

With respect to technologically rooted errors, we try to persuade users that they can very well proceed in their intended exploration through CoRoLa and still obtain desired results. In the future, we intend to design a way to allow users themselves to signal and, in some cases, maybe also make corrections to the texts, for instance by putting up a service that collects errors and correction suggestions from users. Letting users interact with the texts themselves should be monitored, in order to prevent any alterations to the original text beyond mistyping, errors due to formatting or encumbered by technological processing. Examples of such errors are: mistyping => *lima* instead of *limba* (‘language’); formatting: *c lima* instead of *clima* (‘climate’); remained end of line hyphen: *lima- nul istoriei* (approx. ‘the refuge of history’), *proiec- teză lumină* (‘projects light’), *exem- plele pot continua* (‘examples may continue’).

Here are also some usage errors that can be found in CoRoLa.

a. Dragomirescu and Nicolae (2011) show many examples of incorrect uses of words and their meanings. At the time these authors elaborated their collection of examples, CoRoLa did not exist, but we were curious to verify some of their findings. For instance, they discuss excessive proliferation of the passive adjectival participle form *ofertat* (from *a oferta*, ‘to offer, to bid’), with the sense ‘căruia i s-a făcut o ofertă, care a fost invitat’ (‘to whom an offer was made, who was invited’) (Dragomirescu and Nicolae 2011: 62-64). The problem with this passive adjectival participle is of a syntactic nature: a passive adjectival participle cannot be formed from a transitive verb, but only from a transitive passive structure, such as *X este ofertat de către Y* (‘X is being made an offer by Y’).

b. *murit* (‘died’) is the participle form of the verb *a muri* (‘to die’) and is one of the few verbs which has no identical adjectival forms (the adjective is *mort* ‘dead’). To verify in CoRoLa if there are contexts in which *murit* comes after a noun, the following query can be used: [drukola/m="msd:n.*"] [drukola/base=muri & drukola/m="msd:vmp.*"]. Browsing the two pages of occurrences fetched by the interface, the user will quickly understand that the great majority are either due to the existence of a punctuation mark between the two items, to a mistyping or to the deletion of a preceding auxiliary. Within the 45 occurrences reported by the interface, we have found only two that express the looked-for phenomenon: *viață murită* (‘*died life’) and *creştinii muriţi în luptă*[19] (‘The Christians died in battle’). The first

[19] From Alexie I Comnenul.

is a passive participle, the latter, a (perhaps intended) ungrammatical adjectival form. The Romanian language allows nevertheless the use of the participle *muriți* as an adjective or even as a noun (through conversion), but only in rare situation, most of which being found in poetic contexts like *căci morți sunt cei muriți*[20]) ('for dead are those that died', a word for word translation; 'for those who pass the grave come back again no more', a poetical translation[21]).

c. Contemporary Romanian language records some contamination of terms, usually involving neologisms, with the use of which many speakers seem to be unfamiliar. Some instances can also be found in CoRoLa:

(3) *interlocuitor* for *interlocutor* 'interlocutor'
(4) *a acces* 'has acceded'

In most of the cases, the verb *a accede* ('to accede') is not used for past events, as mentioned in DOOM (2005).

(5) *a fi + confortabil* 'to be + comfortable'

The adjective *confortabil* has both the meaning 'convenient' – for an object, and 'snug, relaxed, having no worries or problems' – in terms of people and their mental states. However, only an object (room, chair, cloth) can be *confortabil*, not a person, as it is sometimes used.

d. Eminescu writes in his poem, *Pe lângă plopii fără soț*[22] ('Down where the lonely poplars grow'[23]), *Azi nici măcar îmi pare rău* ('But now I very little care'). The poet consciously elides the double negation [nici (măcar) nu V]. To find out if the lack of double negation is common in Romanian, one possible search in CoRoLa could be [drukola/base="nici.*"][drukola/base!=nu]*[drukola/pos=verb], which brings very few similar contexts.

4.3 New meanings

A number of words in Romanian acquire new meanings with pejorative senses. For instance:

– *mutră* (slang, 'face, facial expression') – usually in plural, as in the idiomatic expression *a face mutre* ('to make faces'), has a depreciative connotation;

– *japiță* – a word missing in some dictionaries, while in others it appears in a strictly technical sense, 'a piece of wood or bent iron, placed above the prot and forming the pit of the yoke' (DM, 1958). In CoRoLa, this term has no occurrence with this meaning, instead it presents a pejorative use, representing an insult to women, and sometimes, even for men.

– *a altoi* ('to engraft') – the current technical meaning (agr. hort.) is that of 'introducing a branch of a plant into the tissue of another'[24], but nowadays this verb acquired a familiar-ironic meaning, like 'beating, kicking, snapping someone'[25].

[20] From the poem *Împărat şi proletar* (1874), in Eminescu, M.,1966.
[21] Translation by Corneliu M. Popescu (Eminescu, M., 1978).
[22] *Pe lângă plopii fără soț* (1883), in Eminescu, M., 1966.
[23] Translation by Corneliu M. Popescu (Eminescu, M., 1978).
[24] The first written attestation, in 1648, cf. DLR, 2010.
[25] Cf. DL, 1955-1958; DM, 1958.

4.4 Identifying expressions

There are many ways to identify expressions commonly used in contemporary Romanian language by querying CoRoLa. In the following we mention some[26]:

a. Querying for Romanian expressions *se spune* or *se zice* ([drukola/orth=se] [drukola/base=spune]), which usually trigger other expressions, occurring either before or after this sequence. Among those there are phrasal units like *a scoate bani (şi) din piatră seacă* (ad litteram, 'to make money (even) out of a dry stone') and *nicio faptă bună nu rămâne nepedepsită* ('no good deed goes unpunished').

b. Using a dictionary of expressions and searching for the lemma of the title word in that dictionary can bring contexts of occurrence of known expressions, like here:

– *a se abate de la...* ('make a digression from'): se [drukola/base=abate] [drukola/m=pos:adposition];

– *a abdica de la...* ('to give up beliefs, ideas, principles'): [drukola/base=abdica] [drukola/m=pos:adposition].

5. CONCLUSIONS

In this paper we described different ways of querying CoRoLa, mainly from a perspective which is closer to the point of view of linguists, touching, in many cases, specifically, lexicographically rooted interests. We believe there is one aspect that should be made clear. Usually, a lexicographer starts from examples of the use of language when editing an entry. CoRoLa certainly allows such searches, by placing the lemma of the targeted word as a criterion. The multitude of examples retrieved by the interface should then be interpreted by the linguist to dissociate semantic uses, sub-senses, etc. CoRoLa also makes possible a different approach to a dictionary entry. Suppose the linguist knows about some specific uses of a word form, as, for instance, in cases of double plurals, and looks for examples that illustrate her/his intuition. To exemplify, the Romanian word *liman* has two plurals: *limanuri* and *limane*… and the intuition is that *limanuri* is mainly used to denote *port; ţărm*, *mal* ('harbor; shore, haven'), including all figurative senses, while *limane* - for *estuar; lac la ţărmul unei mări*; *lagună* ('estuary; a lake on the seashore formed through alluvial blocking the course of a river; lagoon'). We've only started noticing the use of the plural *limane* with figurative meanings over the last few years. Intuitions of this kind can also be verified in CoRoLa, which has to do with language tendencies and evolution.

Many of these interests for consulting CoRoLa should, however, be well tempered, by correctly taking into consideration the limitations of our corpus: it is not balanced between domains and styles, it includes only the last approximately 75 years of written language use, and it does not represent the spoken language in any way[27]. Therefore, if a search for a word yields no results, this should not trigger the conclusion that that word is not used any more. Not finding one word in this… haystack could have a multitude of

[26] From Ilincan (2015).

[27] Although spoken language is represented in the oral component of CoRoLa (see Tufiş *et al.* 2019, in this volume), through OCQP – the Oral Corpus Query Platform, at http://89.38.230.23/corola_sound_search.

causes. As such, the lexicographer should consider CoRoLa searches as indicative: if found, the fetched contexts can give indications of some of its uses, if not found it should only be taken as a first signal before labelling it "out of use".

The opposite thing could happen when the search produces too many extracts. There is a significant difference between the classical methodology of building a dictionary (by exploiting the collection of citations extracted from the selected bibliography) and the one made possible by a huge corpus, as CoRoLa is. Too many examples offered by CoRoLa could make it hard to decide what could be of interest to be included when exemplifying a definition. In the classical way of producing a dictionary, the primary action, that of selecting examples, had to be consumed before the moment of transcribing paper files in the final dictionary entry. In the former case, the lexicographer had to choose what to retain from a much smaller number than the amount obtained by a totally unsupervised process of collection as that offered by searching a corpus. We believe there are two arguments in response to this observation. The first is that the selection of citations in the classical way (which consumes a lot of time and human resources, and which has been ignored in the notice reported above) is included now in the search time. The result is that the time of selection is drastically reduced. Finally, in order to deal with the multitude of outputs, a linguist can use different filters, from exploiting the information encoded in metadata up to using diversified search filters. Only some of these have been shown in our paper. We believe that the user will invent her/his own criteria to improve the productivity of searches. If, on the contrary, the surprise is what is expected from a corpus, the yet unknown uses, then the linguist has to dedicate some time to go through the multitude of examples brought forth.

We are, of course, aware that CoRoLa will gain in usefulness, as related to lexicographers' use at least, only when its contemporary feature will disappear, dissolved in diachronicity, but in order to achieve this much larger coverage, specific concentrated efforts should be dedicated in quite another project.

REFERENCES

Bański, P., E. Frick, M. Hanl, M. Kupietz, C. Schnober, A. Witt, 2013, "Robust corpus architecture: a new look at virtual collections and data access", in: A. Hardie, R. Love (eds), *Corpus Linguistics 2013 Abstract Book*, Lancaster, UCREL, 23–25.

Barbu, A.M., 2018, "Valence Dictionary For Romanian Language In Printed Version And Xml Format", in: V. Păiș, D. Gîfu, D. Trandabăț, D. Cristea, D. Tufiș (eds), *Proceedings of The 13th International Conference "Linguistic Resources And Tools For Processing The Romanian Language"*, Iași, November 22-23, 101–112.

Bingel, J., N. Diewald, 2015, "KoralQuery – a General Corpus Query Protocol", in: G. Grigonyte, S. Clematide, A. Utka, M. Andrius, M. Volk (eds), *Proceedings of the Workshop on Innovative Corpus Query and Visualization Tools at NODALIDA 2015*, Vilnius, May, 11–13, 1–5.

Christ, O., 1994, "A modular and flexible architecture for an integrated corpus query system", in: F. Kiefer, G. Kiss, J. Pajzs (eds), *Papers in computational lexicography, COMPLEX '94*, Budapest, https://arxiv.org/pdf/cmp-lg/9408005.pdf.

Condrea, I., 2015, "Ordinea cuvintelor şi ordonarea ideilor", https://www.timpul.md.

Cornilescu, A., I. Giurgea, 2013, "The adjective", chapter 7, in: C. Dobrovie-Sorin, I. Giurgea (eds), *A Reference Grammar of Romanian. Volume I: The Noun Phrase*, Amsterdam/Philadelphia, John Benjamins, 355–529.

DEX, 2016, *Dicționarul explicativ al limbii române*, ediția a II-a revăzută şi adăugită, Bucureşti, Editura Univers Enciclopedic Gold.

Diewald, N., E. Margaretha, 2017, "Krill: KorAP search and analysis engine", *Journal for Language Technology and Computational Linguistics* (JLCL), 31, 1, 63–80.
DL, 1955-1958, *Dicționarul limbii române literare contemporane*, Bucureşti, Editura Academiei.
DLR, 2010, *Dicționarul limbii române*, ediție anastatică, Bucureşti, Editura Academiei.
DM, 1958, *Dicționarul limbii române moderne*, Bucureşti, Editura Academiei.
DOOM, 2005, *Dicționarul ortografic, ortoepic şi morphologic al limbii române*, ediția a II-a, revăzută şi adăugită, Bucureşti, Editura Univers Enciclopedic.
Dragomirescu, A., A. Nicolae, 2011, *101 greşeli de lexic şi de semantică*, Bucureşti, Editura Humanitas.
Eminescu, M., 1966, *Poezii*, Bucureşti, Editura pentru Literatură.
Eminescu, M., 1978, *Poems*, translation by Corneliu M. Popescu, Bucureşti, Editura Eminescu.
Gîfu, D., A. Moruz, C. Bolea, A. Bibiri, M. Mitrofan, 2019, "The Methodology of Building Corola", in this volume.
Hanks, P., 2013, "Lexical Analysis. Norms and Exploitations", Cambridge Massachusetts, London, The MIT Press.
Ilincan, V., 2015, *Dicționar de expresii românești în contexte* [DERC] *A–C,* Cluj, Presa Universitară Clujeană.
Kupietz, M., N. Diewald, B. Trawiński, R. Cosma, D. Cristea, D. Tufiş, T. Váradi, A. Wöllstein, 2018, "Recent Developments in the European Reference Corpus EuReCo", in: S. Granger, M.-A. Lefer, L. Aguiar de Souza Penha Marion (eds), *Book of Abstracts. Using Corpora in Contrastive and Translation Studies Conference (5th edition)*, Louvain-la Neuve, CECL Papers, 1, 101–103.
Kupietz, M., R. Cosma, A. Witt, 2019, "The Drukola Project", in this volume.
Levin, B., 1993, *English Verb Class and Alternations: A Preliminary Investigation*, Chicago, University of Chicago Press.
McDonald, R., J. Nivre, Y. Quirmbach-Brundage, Y. Goldberg, D. Das, K. Ganchev, K. Hall, S. Petrov, H. Zhang, O. Täckström, C. Bedini, N. Bertomeu Castelló, J. Lee, 2013, "Universal Dependency Annotation for Multilingual Parsing", *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, vol. 2, Sofia, 92–97.
Pala, K., 1999, "Semantic Annotation of (Czech) Corpus Texts", in: V. Matousek, P. Mautner, J. Ocelíková, P. Sojka (eds), *Text, Speech and Dialogue. TSD 1999. Lecture Notes in Computer Science*, vol. 1692, Berlin/Heidelberg, Springer, 56–61.
Pană Dindelegan, G., 1974, *Sintaxa transformațională a grupului verbal în limba română*, Bucureşti, Editura Academiei.
Przepiórkowski, A., Z. Krynicki, Ł. Dębowski, M. Woliński, D. Janus, P. Bański, 2004, "A search tool for corpora with positional tagsets and ambiguities", in: M.T. Lino, M.F. Xavier, F. Ferreira, R. Costa, R. Silva (eds), *Proceedings of the Fourth International Conference on Language Resources and Evaluation* (LREC 2004), Lisbon, Portugal, 1235–1238.
Tesnière, L., 1959, *Éléments de syntaxe structural*, Paris, Klincksieck.
Tufiş D., V. Barbu Mititelu, E. Irimia, V. Păiş, R. Ion, N. Diewald, M. Mitrofan, M. Onofrei, 2019, "Little Strokes Fell Great Oaks. Creating CoRoLa, the Reference Corpus for Contemporary Romanian", in this volume.