Dealing with specialised co-text in text mining: Verbal terminological collocations

Margarida Ramos[(1) (2)], Rute Costa[(1)], Christophe Roche [(2) (1)]

(1) NOVA CLUNL, Centro de Linguística da Universidade NOVA de Lisboa Avenida de Berna 26-C, 1069-061 Lisboa – Portugal

(2) Condillac Group – Listic Lab. Université Savoie Mont-Blanc Campus Scientifique, 73 376 Le Bourget du Lac cedex – France

mvramos@fcsh.unl.pt, rute.costa@fcsh.unl.pt, christophe.roche@univ-savoie.fr

Abstract. The aim of this paper is to organise lexical and conceptual knowledge by analysing a domain-specific corpus. The domain we focus on is the cork industry. Through the analysis of the corpus, we have found that some common verbs in Portuguese, such as "choose" and "separate" acquire a specialised value in the field under study. This was the starting point for the analysis of the terminological collocations where verbs are the core constituents. For the analysis of these verbal terminological collocations, we used natural language processing techniques, building simple and complex CQL structures with RegEx. The outcome of this analysis permits us to introduce a distinction between polylexical terms and terminological collocations. The terminological collocation is a reality of great relevance in specialised discourse, but unlike terms, it is not defined by conceptual criteria, but by morphological and syntactic criteria.

1. Introduction

The aim of this paper is to organise lexical and conceptual knowledge by analysing a domain corpus. The domain we shall focus on is the cork industry. Our work develops around a number of activities that are divided into 3 subsectors based on cork-related tasks: (1) the preparation of cork; (2) its transformation; and (3) the agglomeration of cork products. The texts that make up the corpus report on these activities reflecting the actions required to carry out the specific tasks on each of the sub-sectors identified above. By analysing the corpus, we have found that the verbs "escolher" [choose] and "separar" [separate] acquire a specific specialised value in the field under study and will therefore be the starting point for the analysis of the terminological collocations where verbs are the core constituents.

In order to analyse these verbal terminological collocations (VTC), we have used natural language processing techniques related to text analysis envisaged as one of the tasks involved in text mining:

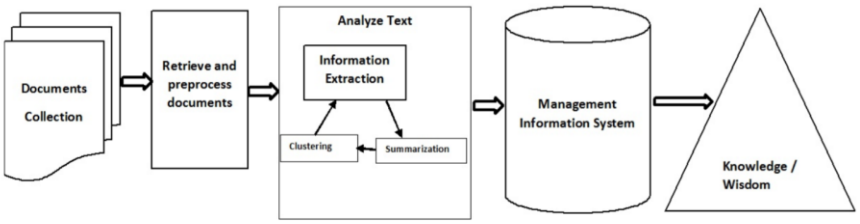

FIG. 1 - *Ramzan, Talib, Muhammad, K. Hanify, Shaeela, Ayeshaz and Fakeeha, Fatima. 2016.*

The texts under analysis are part of a corpus on the topic of cork designed for this purpose. We have used Sketch Engine[1] for text analysis aiming at observing verbal terminological collocations. Our work is based on the previously identified verbs as well as the co-texts in which the verbs occur. By co-text we mean the linguistic sequences that are next to the verbs that were identified for analysis. Using CQL[2], where RegEx (Regular Expressions) are used, we start from the morphosyntactic pattern $[[V_{tr}] + [N]]_{TVC}$, where [V] is either the verb "escolher" [choose] or "separar" [separate], and [N] is a term[3] whose morphosyntactic structure is simple or complex.

[1] https://www.sketchengine.eu/
[2] Corpus Query Language.
[3] a *designation* that represents a *general concept* by linguistic means (ISO 1087:2019)

In Table 1, we present 8 examples found in the corpus corresponding to [N], which is a term – monolexical or polylexical – in the pattern of the structure under analysis:

	Type	Example - PT	EN [literal translation]
1	N	rolha	stopper
2	N+N	cortiça secundeira	secondary cork
3	N+Adj	rolha chanfrada	chamfered stopper
4	N+Prep+N	rolha de cortiça	cork stopper
5	N+N+Adj	cortiça virgem planificada	planned virgin cork
6	N+Adj+Adj	rolha preparada seca	dry prepared stopper
7	N+Adj+Prep+N	rolhas naturais com qualidade	natural stoppers with quality
8	N+Prep+N+Adj	rolhas para utilizações industriais	stoppers for industrial usage

TAB. 1 – *Terminogenic matrix.*

In this paper, we will make some considerations on: (i) verbal terminological collocations; (ii) domain specificities; and (iii) automatic corpus processing using Sketch Engine. Data will be extracted using CQL[4], where RegEx (Regular Expressions) are used, based on a previously prepared terminogenic matrix. The results obtained will be analysed according to the double dimension of Terminology (Costa, 2013).

2. Description of the activities involved in the cork domain

The corpus under analysis is made up of texts produced within the cork industry. As stated in the Introduction, the activity of this industry is divided into 3 subsectors, which include the activities of preparation, transformation, and agglomeration of cork products. The activity of the preparation of cork has to do with slicing [traçamento], stacking [empilhamento], boiling [cozedura], and stabilising [estabelização] the cork. The transformation of cork corresponds to activities that are associated to the manufacture of cork stoppers, which include the production and finishing [acabamento] of cylindrical batons of granulated cork for the manufacture of agglomerated cork stoppers and component-parts of technical stoppers, as well as natural and technical stoppers. The activities of the agglomeration of cork products include the production of materials for the construction industry and for the automobile and aeronautical sectors, among others.

We will focus on the subsector of transformation since the object under study is the production of natural cork stoppers. From the extraction of cork to the final product, several stages are needed, depending on the type of stopper one wants to

[4] Corpus Query Language.

produce. The overall process of the production of the cork stopper is divided into 3 stages, namely debarking [descortiçamento], manufacturing the stopper [fabrico da rolha], and finishing the stopper [acabamento da rolha], where each stage encompasses different processes as illustrated below:

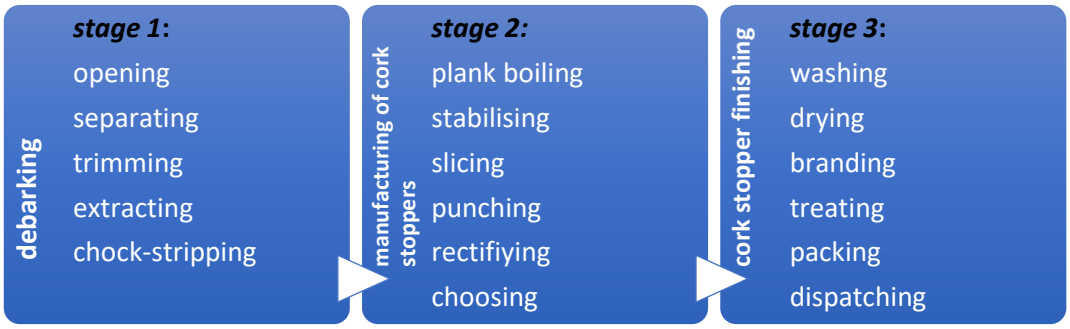

FIG. 2 – *Production of cork stoppers and its different stages (Nunes, 2013).*

The stages of debarking and finishing the stopper are the same for natural and agglomerated cork stoppers. The manufacturing of cork stoppers [fabrico da rolha] shown in Figure 2 relates exclusively to natural cork stoppers, while agglomerated cork stoppers undergo other processes that are included in the agglomeration activity, a process that will not be treated in this context.

The production of natural cork stoppers is one of the transformation activities. This type of stopper is obtained from a thick rectangular-shaped piece of cork named "stripe" [rabanada] through "punching" [brocagem[5]]. Experts explain that stripes are punched [brocadas], and to get those stripes [rabanadas] into a rectangular shape, the planks had to be previously sliced [rabaneadas]. At this stage, after being punched from the stripe, the stopper is only a semi-manufactured product, quite far from being a finished product.

A semi-manufactured natural cork stopper undergoes additional operations until it is a finished product. This is where the finishing process plays its role in the transformation activity. Cork stoppers may be sold with a semi-finished or finished status. The client acquires them (a winery, for instance) either unready or ready to be used, depending on the client's purposes or means to finish the stoppers. Briefly, a semi-finished stopper is a stopper that was submitted to any finishing treatment [tratamento de acabamento] of the finishing process [processo de acabamento], such as rectifying [rectificação], washing [lavação], and subsequent drying [secagem],

[5] Punching is the term used for the manual, semi-automatic or automatic process of perforating the strips of cork with a drill (APCOR: https://www.apcor.pt/en/cork/processing/industrial-path/natural-cork-stoppers/).

except for the final treatment [tratamento final]. At this point, the unready-for-use-stopper is either sold, packed and transported, or continues through the finishing process, until it is ready to be used. To be considered a finished product, the stopper must undergo the final treatments, which are branding [marcação] and/or surface coating treatment.

3. Domain-specific corpus: creation

To attain our terminological goals, we decided to build a domain-specific corpus, i.e., a corpus comprised of texts produced in a specialised context of communication, in which the discourse of a community of experts from a field of interest is reflected. The purpose of the creation of this *corpus* is to analyse the discourses of experts in order to extract information that represents the experts' conceptualisations beyond the verbal expression.

The corpus is comprised of 98 texts written in European Portuguese. These 98 texts were produced by experts belonging to different organisations coming from different areas — scientific, industrial, techno-professional, certifying, regulating, and commercial — and are available online. Within this line, we considered specific criteria for the collection of texts to be included in the corpus, focusing on the communication settings of production/reception, where authorship is of utmost importance for the reliability of the information contained in the texts and the intended outcome of the linguistic analysis. The texts were compiled according to the following criteria:

1. texts produced by and for the scientific community of the domain of cork;
2. texts produced by experts for quasi-experts;
3. texts produced for non-experts.

The rationale behind the inclusion of the third group in the corpus is the fact that these texts are rich in definitional contexts[6] and/or contexts[7] that describe concepts given the different degrees of knowledge of producers and recipients.

[6] By definitional contexts, we mean contexts that are rich in knowledge information permitting the elaboration of definitions.

[7] a piece of text that helps to explain the meaning of a linguistic expression.

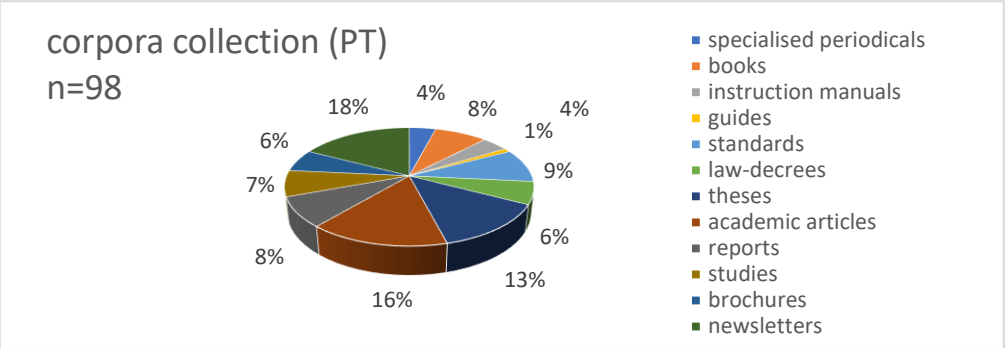

FIG. 3 – *Corpora collection – 98 texts produced following 3 major criteria: expert-expert; expert-quasi-expert; expert-non-experts.*

Following the criteria mentioned above, we obtained a balanced corpus that covers the different levels of specialised discourse.

The communicative setting of the production of the texts was the most significant criterion for the compilation of the corpus to support our terminological purposes. An important aspect is that the linguistic analysis was performed on texts produced by experts for semi-experts or quasi-experts that are commonly technical-explanatory, as well as normative texts, and texts produced for the economic and financial areas (the latter were produced by experts of the domain for experts of governmental institutions). The underlying reason for this option is that these texts contain glossaries and definitions produced by experts; thus, validation of the extracted terms[8] is provided *a priori*. The remaining corpora are used as reference corpora.

4. Collocation

The Anglo-Saxon school (Halliday, 1996; Sinclair, 1996; Benson, 1988, 1997) and the German school (Hausmann, 1989; Heid, 2001) prefer the term collocation to designate sets of units that co-occur in contiguity with a certain frequency in the syntagmatic axis.

In the theoretical framework of Halliday's discourse analysis, the sets of cohesive lexical units deserve attention. For Halliday, cohesion underlies the stabilisation of the constituents that co-occur, and a privileged syntactic-semantic relation can be inferred from that co-occurrence:

[8] The analysis of these glossaries and definitions have been at the core of our terminological work since 2013. Thus, a considerable amount of terms and definitions of the domain has already been compiled.

> Cohesion occurs when the interpretation of some elements in the discourse is dependent on that other. The one presupposes the other, in the sense that it cannot be effectively decoded except by recourse to it (Halliday, 1996, p. 4).

In this theoretical framework, collocations can be analysed from a grammatical or lexical perspective — cohesion, i.e., the essence of collocation as a linguistic entity, is identified by the relationships established between the different lexical units. The lexical units that form a lexical collocation are updated in discourse, because the linguistic system allows for a privileged lexical proximity. This lexical proximity is not always the same:

> There are degrees of proximity in the lexical system, a function of the relative probability with which one word tends to co-occur with another. Secondly, in the text there is relatedness of another kind, relative proximity in the simple sense of the distance separating one item from another, the number of words or clauses or sentences in between (Halliday, 1996, p. 290).

Thus, the density of cohesion is related to how speakers activate the linguistic system and consequently to how they update lexical units, depending on the construction of a discourse or a text.

Sinclair also addresses collocations in their double grammatical and lexical aspects, attributing an essential role to the statistical method for the delimitation of a collocation in order to describe it more adequately:

> Collocation is the occurrence of two or more words within a short space of each other in the text (Sinclair, 1996, p. 170).

Sinclair is particularly concerned with the lexical aspect of collocation, from a lexicographical perspective, focusing on its formal descriptions.

Collocations are thus approached according to lexical and statistical methodologies. Using the concept of distance, understood as the length of the line segment defined by two points, these methodologies allow us to calculate and measure the density between the units that comprise the collocation. The lexical and statistical analyses are, within this scope, complementary tools. The frequency with which each of the units of the collocation is updated in a given syntagmatic order is an indicator of its identification, description, and classification.

Each of the units that make up the collocation may have a different importance and a different value. The unit under observation consists of a node and a collocate, and

any unit may assume the status of node or collocate, depending on the value assigned to each unit.

In 1989, Hausmann defined collocation as:

> […] la combinaison caractéristique de deux mots dans une des structures suivantes: a) substantif + adjectif (épithète); b) substantif + verbe; c) verbe + substantif; d) verbe + adverbe; e) adjectif + adverbe; f) substantif + (prép.) + substantif (Hausmann, 1989, p. 1010).

Hausmann argues that collocation is defined as opposed to free combination and idiomatic expression by the combinatorial restriction of the units that make up the collocation, and by their transparency and their non-cohesion, being apprehended and used as a unit of language and not as a unit of "parole" in the Saussurean sense. The fact that a collocation can assume the status of lexicographic unit implies going from discourse to language at a given moment.

For Hausmann, a collocation is an oriented combination, i.e., the units that comprise it do not have the same status — one of them is the core that is responsible for the privileged lexical relations that it maintains with its local surroundings. For this reason, it is essential to distinguish the base from the collocate ("Kollokator") — which is equivalent to the notions of node and collocate — since its identification is indispensable for linguistic description and the lexicographic treatment of the collocation, as well as to learn it and subsequently assimilate it. These two elements that constitute the collocation have different partnership statuses, and the treatment to which they are subject, depending on whether the emphasis is placed on the base or the collocate, is also different.

A collocation is thus composed of a base with syntactic and semantic autonomy and a collocate, which adds a characteristic to the base, without modifying its identity.

Heid partially follows Hausmann's reasoning. However, his view on collocation is different because it is terminological. Heid explicitly argues that a collocation may correspond to a term with characteristics and properties that are different from those of terms traditionally identified as compound nouns.

Heid (2001) also refers to the polarity of collocations. A collocation is comprised of two lexemes and potential determinants, quantifiers, and prepositions — one of the lexemes is determined and the other is determinant: these notions correspond to the "node" and "collocate" of Sinclair and the "Basis" and "Kollokator" of Hausmann.

Since his approach to collocation is terminological, one of the lexemes must necessarily be a term, and both lexemes may assume this status. For Heid, from a linguistic point of view, collocations are

> […] a phenomenon of lexical combinatories: they involve the lexical, semantic, and syntactic properties of lexical items and their syntagmatic co-occurrence (Heid, 2001:788),

which, like linguistic signs, result from a convention.

When he refers to syntactic properties, Heid establishes a relation between collocation and compound word, since he considers that the choice of the components of a compound is determined from the collocational point of view:

> The choice of the components in such noun groups, like the choice of the components of the compounds, is often collocationally determined: there are clear combinatory preferences, often merely conventional, that in many cases go as far as the complete terminological "fixing" of the compounds and noun groups (Heid, 2001, p. 791).

Heid assumes the difficulty of distinguishing collocation from composition based on purely linguistic criteria, knowing that, from the theoretical point of view, it is a very thin line. From a terminological perspective, such a distinction does not prove to be very operational, since in Terminology it is the designation that is at the basis of the identification of the linguistic reality, regardless of the label that is attributed to it:

> From a terminological point of view, we may be more interested in whether the combination of term and collocate can be seen as the denomination of a new concept in its own right (Heid, 2001, p. 791).

With respect to semantic properties, Heid draws on Mel'čuk's theory (1998). Mel'čuk argues that speakers, in the full use of "parole", use collocations to express generic meaning, and accordingly describes collocations starting from the lexical functions that allow him to account for this generic character and which Heid expresses as follows:

> In lexicography, examples of collocations are usually treated in terms of a given collocate with a given base being arbitrary phenomenon that must be memorized (Heid, 2001, p. 793).

That phenomenon also occurs in specialised language, the difference being that the choice of collocate is usually the result of convention and not of free will.

5. Terminological collocation

In terminology, we are interested in introducing a distinction between polylexical terms and terminological collocations. This distinction is fundamental in terminology, because from our point of view these two lexical phenomena should not be confused. In fact, with ISO 1087: 2019, the term is a *designation* that represents a *general concept* by linguistic means. We would like to stress that from a morphological point of view the term can be monolexical or polylexical.

The terminological collocation is a reality of high relevance in specialised discourse, but unlike terms, it is not defined by conceptual criteria, but by morphological and syntactic criteria, in which a constituent X occurs in a syntagmatic axis in a privileged and frequent way with a constituent Z, and the selection and lexical order is shared by a community of experts. The use of terminological collocations is often evidence of an individual's social and anthropological belonging to a community.

Thus, constituent X corresponds to [V] in our matrix; constituent Z to [N]. In this paper, [V] corresponds to a common verb that, in the cork processing industry, has acquired a specialised value and governs a noun phrase in which [N] is a specialised term, such as:

1. [[separar]$_{V}$ [as [cortiças]$_{N}$] $_{SN}$] $_{VTC}$
2. [[[escolher]$_{V}$ [as [rolhas de cortiça] $_{N}$]$_{SN}$] $_{VTC}$

The collocation is characterised by being composed of a set of elements, where one of them exerts a morphosyntactic and/or semantic attraction over the other constituents that make up the collocation. In the case of a terminological collocation, one of its constituents is a term that in a syntagmatic context attracts another constituent, which may be terminological or not — the whole of that morphosyntactic construction is a non-term. "On considère un non-terme toute combinatoire lexicale qui, d'un point de vue morphosyntaxique, peut se confondre avec un terme – désignation verbale de concept – mais qui n'en est pas un, car cette combinatoire lexicale est non désignative." (Costa 2017).

6. Corpus processing: Sketch Engine

In the scope of this research, we have used Sketch Engine to compile, annotate, and query the corpus employing CQL, where RegEx[9] are used. Sketch Engine has an incorporated tagger for Portuguese called FreeLing, which we have used for the queries.

Considering the 98 documents, we have obtained the following quantitative data:

	Frequency
Tokens	1,712,652
Words	1,217,968
Sentences	48,031

TAB. 2 – *Quantitative data regarding the analysed corpus.*

From the observation of the words identified above, we have seen that the most frequent forms that correspond to terms in the domain under analysis are "cortiça" [cork] and "rolha" [stopper].

Forms	**Frequency**	**Percentage per million**
cortiça	16,127	9,416.40
rolha	7,446	4,347.60

TAB. 3 – *Quantitative data regarding "cortiça" and "rolha".*

Considering the frequency of these two terms and consequently the importance they have in the domain under analysis, we shall look at their behaviour in texts.

The FreeLing tagger has certain limitations, as does Sketch Engine itself. FreeLing cannot distinguish between adjectives and past participles, which is, as regards terminological work, highly limiting as observed in Costa (2001): in terms of probability, in Portuguese, the past participle is not usually part of the morphosyntactic structure of a term, while an adjective can be. This study was performed in the domain of Remote Sensing (Costa, 2001), where that characteristic was observed in the usage of the adjective "colorido" [colourful] as opposed to the usage of the past participle "colorido" [coloured]. This fact causes some noise in the results obtained from CQL queries. On the other hand, Sketch Engine does not allow

[9] a regular expression is a compact way of describing complex patterns in texts. You can use them to search for patterns and, once found, to modify the patterns in complex ways. They can also be used to launch programmatic actions that depend on pattern.
http://gnosis.cx/publish/programming/regular_expressions.html

semantic tagging, which would be a definite plus to retain certain types of forms while rejecting others thus contributing to the reduction of noise in the results obtained.

6.1. Querying the corpus using CQL

We intend to identify verbal terminological collocations (VTC) whose base is the verb "escolher" or "separar" followed by a specialised term used in the domain under analysis, and whose pattern follows a recurring morphosyntactic pattern in Portuguese: [[V] + [N = mono or polylexical term]]$_{VTC.}$ The structure of the term analysed here corresponds to a noun phrase, which, for example, can have the following behaviour:

[N+N] Term
[N+Prep+N] Term
[N+Adj] Term

In the analysed corpus, we have identified structures such as *separar rolhas com defeitos* [V+N= polylexical term] [separating stoppers with defects] or *estabilizar a cortiça cozida* [V+N= polylexical term] [stabilizing the boiled cork]. Hence, the first structure shall be segmented as [V +[N+Prep+N]$_{N= Term}$]$_{VTC}$ and the second one as [V + [N+Adj]$_{N=Term}$]$_{VTC}$. "Rolhas com defeitos" [stoppers with defects] and "cortiça cozida" [boiled cork] are terms because they designate concepts.

Based on our linguistic knowledge, we employed CQL using regular expressions to identify fundamental lexical-semantic patterns in verbal terminological collocations.

The tags we have used for the CQL constructs are those adopted by Sketch Engine for the Portuguese language: FreeLing part-of-speech tagset[10], a morphological tagger based on EAGLES[11] proposals.

By using Sketch Engine, we are restricted to the part-of-speech tags available of FreeLing in our queries (CQL A to F – see Table 4 below), such as V; VM; VP; D; A and N. Each of these labels have the following value: V=Verb; VM=Main verb; VP=Past Participle; D=Determiner; A=Adjective; and N=Noun. In addition to these, we also used the "character class" [[:punct:]], a RegEx construct, in order to reflect our wish of "no punctuation" in the results (see CQL F, Table 4).

[10] "A tagset is a list of part-of-speech tags (POS tags for short), i.e. labels used to indicate the part of speech and sometimes also other grammatical categories (case, tense etc.) of each token in a text corpus": https://www.sketchengine.eu/portuguese-freeling-part-of-speech-tagset/?highlight=freeling.

[11] http://www.ilc.cnr.it/EAGLES/browse.html

The decision to use the generic tags V=Verb, N=Noun and A=Adjective in the first queries, instead of specific subtypes such as VM=Main Verb or VP=Past Participle, is a consequence of the limitations of FreeLing. We have noticed amongst the results of our CQL and also in the Word Sketch[12] for "rolha" [stopper], that some linguistic forms are either tagged as A or N, such as "técnica" [technical], or tagged as VP instead of A, such as "cobrilhada" [faulty], among others. Therefore, we decided to construe the first CQL in a somewhat non-linguistic sense, but in such a way that one does not mismatch valuable results.

Below we present six possible examples of CQL that correspond to the following combinations:

	Corpus Query Language
CQL A	[tag="V.*"][tag="D.*"]?"cortiça.*|rolha.*" [tag="V.*|N.*"]
CQL B	[tag="V.*"][tag="D.*"]?"cortiça.*|rolha.*"[]{0,2}[tag="V.*|N.*"]
CQL C	[tag="V.*"][tag="D.*"]?"cortiça.*|rolha.*"[]{0,2}[tag="V.*|N.*"]
CQL D	[tag="V.*" & !(tag="V.P.*")] [tag="D.*"]?"cortiça.*|rolha.*" []{0,2}[tag="V.P.*|N.*"]
CQL E	[tag="VM.*" & !(tag="V.P.*") & !(lemma="ser.*")] [tag="D.*"]?"cortiça.*|rolha.*" [tag="V.P.*|A.*|N.*"]
CQL F	[tag="VM.*" & !(tag="V.P.*") & !(lemma="ser.*")] [tag="D.*"]?"cortiça.*|rolha.*" [word=".*" & word!="[[:punct:]]*"]{0,2} [tag="V.P.*|A.*|N.*"]

TAB. 4 – *CQL queries on the analysed corpus.*

With these six CQL queries we intend to identify the co-texts in which the terms "cortiça" and "rolha" occur, where the obligatory condition is that they are objects of a transitive verb, [V + [N] *polylexical term*] VTC, amounting to all the parts of a VTC. Building on the analysis of the results obtained with CQL A, we fine-tuned CQL queries until we obtained CQL F, which allowed us a finer granularity of the results obtained with CQL queries.

6.2. Example: CQL F (cf. Table 4)

CQL F reads as follows, using Python operators incorporated into Sketch Engine:

> Verb (Main transitive verb EXCEPT Past Participle and EXCEPT lemma of the verb to be) + Determinant (or not) + forms started by *cortiça* or *rolha* + [occurrence of zero to 2 forms (any) EXCEPT punctuation] + Verb (any) ONLY Past Participle) or Adjective (any) or Noun (any).

[12] A summary of a word's behaviour – one of Sketch Engine features.

We then applied 4 filters, whose starting point is a transitive main verb followed by a morphosyntactic sequence where the term "cortiça" occurs.

	Corpus Query Language F	**Frequency**
1	tag="VM.*" & !(tag="V.P.*") & !(lemma="ser.*")] [tag="D.*"]"cortiça.*" [word=".*" & word!="[[:punct:]]*"]{0,2} [tag="V.P.|A.*"]	68
2	[tag="VM.*" & !(tag="V.P.*") & !(lemma="ser.*")] [tag="D.*"]"cortiça.*" [word=".*" & word!="[[:punct:]]*"]{0,2} [tag="N.*"]	123
3	[tag="VM.*" & !(tag="V.P.*") & !(lemma="ser.*")] "cortiça.*" [word=".*" & word!="[[:punct:]]*"]{0,2} [tag="V.P.*|A.*"]	58
4	[tag="VM.*" & !(tag="V.P.*") & !(lemma="ser.*")] "cortiça.*" [word=".*" & word!="[[:punct:]]*"]{0,2} [tag="N.*"]	73

TAB. 5 – *Extension of CQL F for the co-text of the term "cortiça".*

After applying all the rules, we obtained the following results:

CQL F	**Good examples**			**Examples on which to decide**		**Bad examples**	
1	separar a cortiça com verde	separando-se a cortiça virgem		golpeia-se a cortiça no sentido vertical	identificar a cortiça com verde	extrair a cortiça com diversos	retirar a cortiça em grandes
2	estabilizar a cortiça após o descortiçamento	Preparar a cortiça para a transformação		identificar a cortiça com verde fresco	retirar a cortiça em grandes pranchas	Estabilizar a cortiça de forma	promover a cortiça e o reforço
3	--	produzir cortiça para utilizações industriais		produzir cortiça de forma sustentável	--	usando cortiça como componentes	
4	extrair cortiça dos ramos	retirar cortiça com maior calibre	tirar cortiça de um sobreiro	--	--	encontrar cortiça com o calibre	tinham cortiça virgem nesse momento

TAB. 6 – *Results obtained applying CQL F and its extensions.*

The decision whether examples are “good examples”, “examples on which to decide” or “bad examples” is based on the knowledge that the authors have of the domain and the corpus, as well as on the linguistic knowledge they have regarding word and term formation and the formation of collocations. It should be noted that all the examples shown in Table 6 have not been submitted to expert validation yet. Depending on the expert feedback, data may be reorganised, if necessary.

We repeated the same exercise replacing “cortiça” [cork] with “rolha” [stopper].

Following this CQL, we applied four filters whose starting point is a transitive main verb followed by a morphosyntactic sequence where the term “cortiça” [cork] occurs:

	Corpus Query Language F	**Frequency**
1	[tag="VM.*" & !(tag="V.P.*") & !(lemma="ser.*")] [tag="D.*"]"rolha.*" [word=".*" & word!="[[:punct:]]*"]{0,2} [tag="V.P.*|A.*"]	142
2	[tag="VM.*" & !(tag="V.P.*") & !(lemma="ser.*")] [tag="D.*"]"rolha.*" [word=".*" & word!="[[:punct:]]*"]{0,2} [tag="N.*"]	286
3	[tag="VM.*" & !(tag="V.P.*") & !(lemma="ser.*")]"rolha.*" [word=".*" & word!="[[:punct:]]*"]{0,2} [tag="V.P.*|A.*"]	61
4	[tag="VM.*" & !(tag="V.P.*") & !(lemma="ser.*")]"rolha.*" [word=".*" & word!="[[:punct:]]*"]{0,2} [tag="N.*"]	71

TAB. 7 – *Extension of CQL F for the co-text of the term “rolha”* [*stopper*].

Applying the rules, we obtained the following results:

CQLF	**Good examples**			**Examples on which to decide**		**Bad examples**	
1	separar as rolhas acabadas	separar as rolhas mal coladas	segregar as rolhas mal colmatadas	--	rastrear as rolhas prontas	classifiquem as rolhas em classes visuais	correspondem a rolhas potenciais foram designadas
2	brocar uma rolha de 24 mm	lavarmos uma rolha de cortiça	--	retirar as rolhas das rabanadas	mergulhar as rolhas com agitação	lavar as rolhas também após colmatagem	transformam as rolhas em bruto
3	--	brocar rolhas naturais	--	produzindo rolhas cilíndricas	fazer rolhas técnicas	contendo rolhas acabadas prontas	separam-se as rolhas por classe
4	rectificar rolhas de cortiça natural	brocar rolhas naturais com qualidade	desinfectar rolhas de cortiça	--	exportavam rolhas dos portos	ver rolhas técnicas	abrangeu rolhas de diversas qualidades

TAB. 8 – *Results obtained applying CQL F and its extensions.*

According to the examples sampled, we can associate V + term "cortiça" [cork] or "rolha" [stopper] and obtain satisfactory results to identify verbal terminological collocations.

In the following lines, we describe some procedures and operations, which are designated by the terms presented in FIG. 2. The attempt is to demonstrate how these terms occur in specialised texts by means of their morphosyntactic behaviour. In a domain of handicraft and industrial activities, verbs are at the core of the coinage of terms, which accounts for our terminological interest on this subject.

After analysing our corpus, we have identified different verbs to designate the same process. For instance, in the manufacture of the stopper stage, the verbs "escolher" [choose] and "selecionar" [select] are synonyms in the corpus. On the other hand, we also noticed that "escolher" [choose] is replaced by "separar" [separate], both used as synonyms in discourse, although they are only quasi-synonyms since the acts of choosing and separating do not correspond to the same action[13]. We can see that this discursive option introduces ambiguity in the description of the activities of the domain since the verb "separar" [separate] is also used in the debarking stage.

1.CQL: [lemma="separar.*"] []{0,2} "cortiça.*"			
left context	**KWIC**	right context	freq
componente principal	**separa** as cortiças	segundo as	1
com o intuito de	**separar** porções de cortiça	em grupos,	1
a variável que	**separa** as cortiças	é área mínima	1
ou seja, é	**separada** a cortiça	que não possui	1
as pranchas,	**separar** a cortiça	com verde e	2
,segregadas,	**separadas** da restante cortiça	destinada à	7
poderá ser	**separada** da cortiça	delgada	1
da extracção,	**separando**-se a cortiça	virgem e bocados	1
2.CQL: "cortiça.*" [] {0,2} [lemma="separar.*"]			
left context	**KWIC**	right context	freq
os grãos de	cortiça, previamente **separados**	em gamas de	1
colhidas amostras de	cortiça e **separadas**	as amostras rolháveis	1
da cozedura, as	cortiças são **separadas**	em fardos de acordo	1
conjunto de pranchas de	cortiça preparada **separadas**	em diferentes classes	3
As pranchas de	cortiça devem ser **separadas**	do solo por	3
impacto dos martelos na	cortiça (**separam**	a lenha da cortiça)	1
3.CQL: [lemma="seleccionar.*"] []{0,2} "cortiça.*"			
left context	**KWIC**	right context	freq
cortiça cuidadosamente	**selecionados** aglutinados com cortiça	. São diversas as	1
4.CQL: "cortiça.*" [] {0,2} [lemma="selecc?ionar.*"]			
left context	**KWIC**	right context	freq
o crescimento da	cortiça **seleccionam-se**	da bibliografia	1
para o crescimento	cortiça foi **seleccionado**	tendo em conta	1
conjunto de dez	cortiças **seleccionadas**	na oficina do	1

[13] "separar" [separate]: fazer a disjunção de [divide]; "escolher" [choose]: manifestar preferência por [show preference for] (Dicionário do Houaiss, 2003).

as pranchas de	cortiças são **seleccionadas**	de acordo com a	1
com granulados de	cortiça cuidadosamente **seleccionados**	e aglutinados com borracha	1
O pavimento em	cortiça foi **seleccionado**	para responder às	1
com granulados de	cortiça cuidadosamente **seleccionados**	em que o látex	1
com granulados de	cortiça cuidadosamente **seleccionados**	aglutinados com cortiça	1
dois ou três discos de	cortiça natural **seleccionada**	As rolhas aglomeradas	2
Passado este período ,	cortiça é então **seleccionada**	nomeadamente no que	2
5.CQL: [lemma="escolher.*"][]{0,2}"cortiça.*"			
left context	**KWIC**	right context	freq
preparadora. Esta indústria	**escolhe** as cortiças	empilhadas de acordo	1
Herdade de Espirra foram	**escolhidas** as cortiças	em função do calibre	1
monge beneditino,	**escolheu** as rolhas de cortiça	para vedar o seu famoso	1
6.CQL: "cortiça.*"[]{0,2}[lemma="escolher.*"]			
left context	**KWIC**	right context	freq
as pranchas de	cortiça amadia, **escolhe**	novamente por qualidades	1
ou seja, das	cortiças **escolhidas**	e classificadas “(	1
7.CQL: [lemma= "separar.* "] []{0,2} "rolha.*"			
left context	**KWIC**	right context	freq
operação destinada a	**separar** as rolhas	acabadas em classes	3
componente principal	**separa** as rolhas	que apresentam valores	1
, ou seja,	**separam**-se as rolhas	por classe e com defeito	1
Operação destinada a	**separar** as rolhas	em determinado número	20
encontrar-se fisicamente	**separadas** das rolhas	e dos discos,	1
obrigatórias: 5.3.1	**Separar** as rolhas	em função das referências	1
devem estar fisicamente	**separadas** das rolhas	não lavadas, quer	1
que se destina a	**separar** as rolhas	com defeitos de colagem	3
4.2 Objectivo :	**Separar** as rolhas	mal coladas	2
que se destina a	**separar** as rolhas	com defeitos 3.2	4
que consiste em	**separar** as rolhas	ou discos em várias categorias	1
eu aspecto visual e /ou	**separar** as rolhas	com defeitos 2.3	4
5.2 Objectivo :	**Separar** as rolhas	mal coladas.	1
8.CQL: "rolha.*" [] {0,2} [lemma="separar.*"]			
left context	**KWIC**	right context	freq
programadas e as	rolhas são **separadas**	, com um mecanismo	1
espumantes . 922 . Estas	rolhas estão geralmente **separadas**	em classes "Extra",	1
imperfeições que as	rolhas apresentem, **separando**-as	concomitantemente, em classes	1
9.CQL: [lemma="seleccionar.*"] []{0,2} "rolha.*"			
left context	**KWIC**	right context	freq
de cortiça natural	**seleccionada** . As rolhas	aglomeradas são inteiramente	2
10.CQL: "rolha.*" [] {0,2} [lemma="selecc?ionar.*"]			
left context	**KWIC**	right context	freq
Selecção: processo no qual as	rolhas são **seleccionadas**	de acordo com a sua qualidade	1
O comprimento da	rolha **seleccionada**	deve estar de acordo	1
11.CQL: [lemma="escolher.*"][]{0,2}"rolha.*"			
left context	**KWIC**	right context	freq
monge beneditino,	**escolheu** rolhas	de cortiça para vedar	1

normal) , devem-se	**escolher** rolhas	com um diâmetro superior	2
12.CQL: "rolha.*"[]{0,2}[lemma="escolher.*"]			
left context	KWIC	right context	freq
qualidade associado. As	rolhas depois de **escolhidas**	separadas seguem para	1
- Escolha visual As	rolhas são **escolhidas**	em máquinas electrónicas	1

TAB. *9 – Verbs "separar"[separate]; "seleccionar" [select]; and "escolher" [choose] in co-text with "cortiça" [cork] or "rolha" [stopper]. KWIC were drawn from the cork corpus with CQL queries.*

Table 9 contains the verbs "separar" [separate] and "escolher" [choose] in co-occurrence with "cortiça" [cork], obtained from the cork corpus using CQL interrogations. The verbs that are the starting point in the CQL are highlighted, either on the left or right-side of the key word in context (KWIC). We can observe that both "cortiça" [cork] and "rolha" [stopper] widely co-occur with several inflexions of the verb "separar" [separate] (e.g., *separa as cortiças; separada a cortiça; separar a cortiça / separam-se as rolhas; separar as rolhas; separadas das rolhas*). However, while "cortiça" [cork] has a high co-occurrence with "seleccionar" [select], "rolha" [stopper] has a very low co-occurrence with this verb, as seen on CQL 3. and 4. Vs. CQL 9. and 10. Finally, the verb "escolher" [choose] has a shred of minor evidence for both "cortiça" [cork] and "rolha" [stopper], as shown on CQL 5.; 6.; 11; and 12.

7. Conclusions

The purpose of this research was to prove that verbal terminological collocations are linguistic structures that, together with polylexical terms, play a fundamental role in expert discourse. However, they perform different functions: although they may have morphosyntactic and lexical structures that are actually the same or similar, polylexical terms and terminological collocations are distinguished by the criteria underlying the analysis: terms are governed primarily by conceptual criteria and collocations by morphosyntactic criteria.

The analysis we have carried out in this paper aims to demonstrate how morphosyntactic analysis is complementary to a more concept-focused analysis, allowing us to obtain information that can feed different terminological resources (dictionaries, ontologies, ...).

In the domain of the cork industry, common verbs in Portuguese acquire specific meaning when occurring in co-text with terms; an evidence observed through the analysis of the recursive morphosyntactic constructions found in the corpus. These structures underpin our distinction of a polylexical term from a verbal terminological collocation.

This paper had three purposes that the authors believe to be fundamental for the terminological work:

1. Associating domain knowledge and the linguistic analysis of how texts work;
2. Based on that knowledge, creating local grammars from the analysis of co-texts, in this case, for transitive verbs;
3. Using text mining tools to increase knowledge on the behaviour of the combinations under analysis;
4. Including data validation criteria.

Automatic language processing tools have their limitations. Sketch Engine is no different. Some of the bad results obtained are originated by FreeLing limitations, which forces the user to be somewhat creative in order to capture any meaningful silence and/or eliminate noise. Using this methodology, text processing is still an overly labour-intensive and time-consuming task.

The work that has been carried out at NOVA CLUNL since 2001 is now being updated so it contains semantic information that will increase the quality of the data obtained.

Acknowledgements: Research financed by Portuguese National Funding through the FCT – Fundação para a Ciência e Tecnologia as part of the project Centro de Linguística da Universidade NOVA de Lisboa – UID/LIN/03213/2020 and the PhD program in Linguistic KRUse - Knowledge, Representation & Use, CLUNL - Faculty of Social Sciences and Humanities, Universidade NOVA de Lisboa - PB/BD/113972/2015.

References

Benson, Morton, Benson, Evelyn and Ilson, Robert. 1988. "The BBI Combinatory Dictionary of English. A Guide to Word Combinations". In *Revue belge de philologie et d'histoire*, Vol. 66 No. 3, 709-710. Langues et littératures modernes - Moderne taal- en letterkunde.

Benson, Morton, Benson, Evelyn and Ilson, Robert. 1997. "The BBI Dictionary of English Word Combinations.", VII - XXXIX. Amsterdam, Philadelphia: John Benjamins

Costa, Rute. 2017. "Les collocations terminologiques." Provas de agregação, Lexicologia, Lexicografia, Terminologia. Lisbon: FCSH UNL.

Costa Rute. 2013. "Terminology and Specialised Lexicography: two complementary domains". In *Lexicographica*. *International Annual of Lexicography*, Vol. 29 No. 1, edited by Gouws, Rufus Hjalmar / Heid, Ulrich / Schierholz, Stefan J. / Schweickard, Wolfgand / Wiegand, Heribert Ernst. Berlin, New York: De Gruyter.

Costa, Rute. 2001. "Pressupostos teóricos e metodológicos para a extracção automática de unidades terminológicas multilexémicas". PhD dissert., Lisbon: FCSH UNL.

Halliday, M. A. K..1991. "Corpus Studies and Probabilistic Grammar". In *English Corpus Linguistic, Studies in Honour of Jan Svartvik,* edited by Karin Aijmer & Bengt Altenberg, 30 - 43. London, New York: Longman.

Hausmann, Franz Josef. 1989. "Le dictionnaire des collocations". In *Wörterbucher, Ein internationales Handbuch fur Lexicographie*, edited by Franz Josef Hausmann, Oskar Reichmann, Herbert Ernst Wiegand, Ladislav Zgusta, 1010 - 1019. Berlin, New York: Walter de Gruyter.

Heid, Ulrich. 2001. "Collocations in Sublanguage Texts: Extraction from Corpora." In *Handbook of Terminology Management. Application-Oriented terminology Management*, Vol. 2, compiled by Sue Ellen Wright and Gerhard Budin, 788 - 808. Amsterdam, Philadelphia: John Benjamins Publishing Company.

ISO 1087-1. 2019. "Terminology work and terminology science — Vocabulary". Genève: Organisation Internationale de Normalisation.

Mel'cuk UK, Igor A..1998. "Collocations and Lexical Functions, Phraseology, Theory, Analysis, and Applications", edited by A. P. Cowie, 23 – 54. Oxford: Oxford University Press.

Nunes, Paulo. 2013. "Análise do fluxo de processo industrial e do respetivo plano de inspeção e ensaios." Ma dissert., Porto: FEUP Universidade do Porto.

Ramos, Margarida and Costa, Rute. 2018. “Semantic Analyses of Texts for Eliciting and Representing Concepts: the TermCork Project.” In *Actes de la dixième conférence TOTh 2016, 9-10 June.* Chambéry: Institut Porphyre, Savoie et Connaissance.

Ramzan, Talib ; Muhammad, K. Hanify ; Shaeela, Ayeshaz & Fakeeha, Fatima. 2016. “Text Mining: Techniques, Applications and Issues”. In *(IJACSA) International Journal of Advanced Computer Science and Applications,* Vol. 7 No. 11, 414 – 418. Available at https://thesai.org/Publications/IJACSA.

Silva, Raquel & Costa, Rute. 2019. “Accéder aux connaissances des experts par l’entremise de la médiation en terminologie”. In *L’essentiel de la médiation. Le regard des sciences humaines et sociales,* edited by Michele De Gioia and Mario Marcon, 105 – 121. Bruxelles, Bern, Berlin, New York, Oxford, Wien: P.I.E. Peter Lang

Sinclair, John: Ball, J. 1996. “EAGLES: Preliminary Recommendations on Text Typology (EAG - TCWG - TTYP/P).”, Version of June, pp. 71. Available at http://www.ilc.pi.cnr.it.

Résumé

Le but de cet article est d'organiser les connaissances lexicales et conceptuelles en analysant un corpus spécifique à un domaine. Le domaine sur lequel nous nous concentrons est l'industrie du liège. Grâce à l'analyse du corpus, nous avons constaté que certains verbes communs en portugais, tels que « choisir » et « séparer » acquièrent une valeur spécialisée dans le domaine à l'étude. Ce fut le point de départ de l'analyse des collocations terminologiques où les verbes sont les constituants centraux, dans la perspective de la double dimension de la terminologie. Pour l'analyse de ces collocations terminologiques verbales, nous avons utilisé des techniques de traitement du langage naturel dans lesquelles des structures CQL simples à plus complexes ont été construites avec REGEX. Le résultat de cette analyse nous permet d'introduire une distinction entre termes polylexicaux et collocations terminologiques. La collocation terminologique est une réalité d'une grande pertinence dans le discours spécialisé, mais contrairement aux termes, elle n'est pas définie par des critères conceptuels, mais par des critères morphologiques et syntaxiques.