Extracting knowledge-rich information from definitions. A corpus-based approach to building a conceptual-based terminological resource

Margarida Ramos, Rute Costa

Extracting knowledge-rich information from definitions. A corpus-based approach to building a conceptual-based terminological resource

Margarida Ramos [1], Rute Costa [1]

[1] *NOVA CLUNL, Centro de Linguística da Universidade NOVA de Lisboa, Avenida de Berna 26-C, 1069-061 Lisboa, Portugal*

Abstract

This paper aims to describe a text-mining approach on a domain corpus (cork) within the theoretical framework of the dual dimension of terminology to create a terminological dictionary and correlate it with an ontology. We will make some considerations on (i) domain specificities; (ii) lexical markers; (iii) automatic corpus processing using Sketch Engine; (iv) representation of lexical networks using CmapTools; and (v) representation of the concept system using Protégé. The goal of the ontology is to logically support the coherence and quality of the natural language definitions contained in the terminological resource.

Keywords

terminology; definition; domain-ontology; knowledge-rich information; domain corpus; terminological dictionary.

1. Introduction

This paper aims to demonstrate a method for developing a terminological dictionary based on a domain ontology. To this end, we will describe the methods used to capture specialised lexical and conceptual knowledge from the corpus and use it to develop a dedicated ontology. The terminological resource will consist of a linguistic description of the specialised concepts, based on the formal definitions of the concepts that make up the cork ontology, the OntoCork [11].

The method used in this paper is corpus-driven. The corpus was compiled based on rigorous criteria specific to terminological work [10], where the specialised context of text production is a key-element. In this sense, the corpus is composed of technical explanatory and normative (standards) texts. For corpus analysis, we used Sketch Engine[2] to find and systematise lexical-semantic relationships. During the corpus analysis process, we found two types of relevant knowledge-rich information [6]: definitions and definitional contexts. Definitions are one of the components of the glossary's microstructure that can be found at the end of the normative texts. The purpose of these definitions is to achieve a consensus among the members of the cork community. On the other hand, the definitional contexts are integral parts of the texts and have relevant specialised lexical-semantic markers in their structure.

Our method encompasses two stages:

(i) From the linguistic analysis of the lexical markers, and the corresponding lexical-semantic relations observed between the terms, we systematise the results into lexical maps using CmapTools[3].

(ii) Based on the previous stage, we proceed to the conceptual analysis and subsequent formal representation. The conceptual analysis grounds the identification of conceptual relations obtained by interpreting the lexical-semantic relations observed between two terms. To infer conceptual relations –

2nd International Conference on "Multilingual Digital Terminology Today. Design, Representation Formats and Management Systems" (MDTT 2023), June 29–30, 2023, Lisbon, Portugal

EMAIL: mvramos@fcsh.unl.pt (M. Ramos); rute.costa@fcsh.unl.pt (R. Costa).

ORCID: 0000-0001-7209-3806 (M. Ramos); 0000-0002-3452-7228 (R. Costa)

[2] https://www.sketchengine.eu/

[3] https://cmap.ihmc.us/

such as the associative type – and to identify characteristics that will help us in the process of building concept systems, we resort to deductive mechanisms employing the Aristotelian formula: X=Y+DC to build OntoCork.

2. Domain corpus: cork

The Cork Corpus was built up from texts produced within the cork industry. The internal and external criteria [1],[8] used to build our specific-domain corpus are systematised in Table 1.

Error! Reference source not found. **1**
Internal and external criteria of the cork corpus

Criteria	Purpose/description
Degree of specialisation	Produced by experts and semi-experts
Source validation	Entities recognised as an authority
Type	Technical-explanatory; normative
Content adequacy	On cork/Cork stopper
Synchronism (≤ 10 years)	Given the fast evolution of technology

The corpus comprises 98 texts written in European Portuguese (see Figure 1).

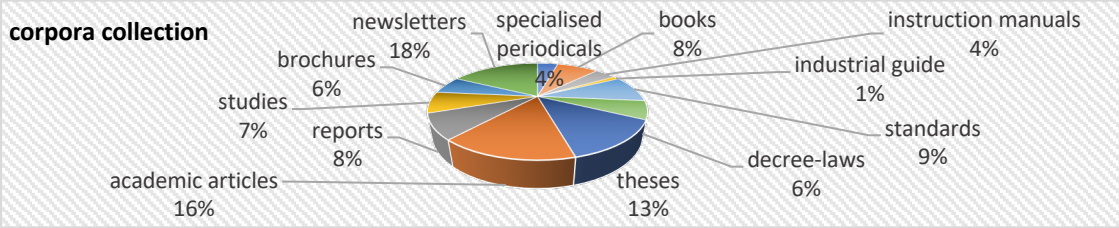


Figure 1: Corpora collection

These texts were produced by experts from different organisations and in different domains related to the cork industry. The texts were collected according to the following criteria: (1) texts produced by and for the scientific community in the domain of cork; (2) texts produced by experts for quasi-experts; and (3) texts produced by experts for non-experts.

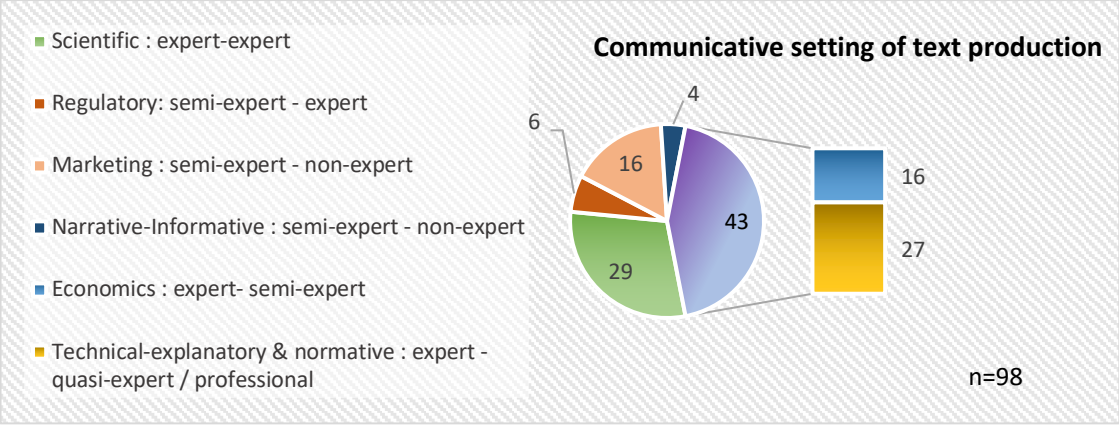
3. Terminological data extraction

Considering the 98 documents of the corpus, we have obtained the quantitative data shown in Table 2.

Error! Reference source not found. **2**
Quantitative data of the corpus

	Total number
Tokens	1,712,652
Words	1,217,968
Sentences	48,031

For the corpus exploration and linguistic data analysis, we mainly focused on 43 texts produced in two communicative settings, namely (i) expert–semi-expert, and (ii) expert–quasi-experts/professionals, while the remaining 55 texts were used as a reference corpus [2] so that we could compare a given terminological data extraction (see Figure 2). The corpus was processed using Sketch Engine, with which we compiled, annotated, and queried the corpus employing advanced searches in Corpus Query Language (CQL) format, where regular expressions (regex) are applied.



Error! Reference source not found. **2:** Subcorpus under focus (43 texts) based on the communicative setting of text production

Among the results presented in Table 2, the most frequent terms are "cortiça" [cork][4] and "rolha" [stopper]. Given the high frequency of these two terms, we analysed the contexts in which they occur in the subcorpus (43 texts) using the Word Sketch function as a first option, with which we identified some candidate terms such as "ROLHA COLMATADA" [colmated stopper] (in capital letters). We then moved on to simple queries (concordances) to search for polylexical terms containing adjectives in their pattern. Once the most common morphosyntactic structures of terms were identified, we decided to improve our search for terms and definitions employing advanced queries, namely through regex, so that we could capture knowledge-rich contexts (KRC) [6], e.g., definitions (definitions found in context) and definitional contexts (contexts explaining what the concept is; thus valuable for understanding and/or elaborating proper definitions).

3.1. Exploring the corpus with text mining methods

Based on the patterns we have identified within definitional contexts, we explored the subcorpus with advanced queries using regexes. For this paper, we will highlight two specific regexes that proved productive in isolating lexical relations between terms, but also in finding definitional contexts where the generic term is expanded in its syntax (see Table 3).

Error! Reference source not found. **3**
Linguistic expressions commonly used by experts

	Definitional contexts (pt)	**Literal translation into English**
(1)	Rolha que foi submetida a um tratamento químico com o objectivo de desinfectar e/ou homogeneizar a cor e/ou branquear.	en: Stopper that was submitted to chemical treatment with the aim of disinfecting and/or homogenising the colour and / or bleaching
(2)	Rolha cuja superfície lateral foi submetida a uma operação de abrasão para a tornar cilíndrica ou diminuir o seu diâmetro.	en: Stopper whose side surface was submitted to an abrasion operation to make it cylindrical or to reduce its diameter.]

The first regex has the following structure: `"rolha"[tag="V.P.*SF"]`, whose formulation aims to match patterns as ONLY forms of "rolha" [stopper] followed by ANY past participle ONLY in the singular and feminine inflection. For the elaboration of this regex, we considered the linguistic expressions used repeatedly by the experts, such as the past participle co-occurring with a term (see Table 4). The outcomes of this query, namely 69 hits, delivered the most productive patterns for identifying lexical markers, such as "x foi submetida a y" [x was submitted to y], as well as terms whose morphosyntactic structures fall under our search patterns, such as [Noun + Past Participle], e.g., "x acabada" [finished X] or "X terminada" [finalised X] where X is a term and Y corresponds to a structure that has proved

[4] Our translation

to be rich in knowledge information, i.e. information provided by the experts that allows us to perceive their conceptualisations [9].

Considering the satisfactory results obtained, we decided to expand its formulation (regex 2): `"rolha"[(tag="D.*"|(tag="S.*")]?[tag="A.*"]?"cortiça"?[]{0,4}"rolha"[]{0,4}[tag="V.P.*SF"]`. In this case, we want to match a context in which the terms "rolha" [stopper] and "cortiça" [cork] may co-occur with either adjectives, past participles, or found duplicated, in addition to the functional forms. Out of the 55 hits matched by this regex, 48 were either a description or a definition.

From the whole set of descriptions or definitions semi-automatically extracted from the Cork Corpus, we decided to select ten (10) definitions for linguistic and conceptual analysis (see Table 4).

Error! Reference source not found. **4**
Ten (10) definitions to organise a typology of cork stoppers

#	10 definitions (literal translations from pt)	10 definitions (pt) extracted from the Cork Corpus
1	stopper Product obtained from natural cork and / or agglomerated cork, consisting of one or more pieces, intended to seal bottles or other containers and to preserve their contents. (5.1 - NORM)	rolha Produto obtido da cortiça natural e / ou de cortiça aglomerada, constituído por uma ou mais peças, destinado a vedar garrafas ou outros recipientes e a preservar o seu conteúdo. (5.1 - NORM)
2	STOPPER piece of cork, usually cylindrical, conical or prismatic quadrangular, sometimes with rounded or chamfered lateral edges, consisting of one or several glued elements and intended to seal the containers or contribute to their water tightness. (7.8 – TECH)	ROLHA peça de cortiça, em geral cilíndrica, troncocónica ou prismática quadrangular, por vezes de arestas laterais boleadas ou chanfradas, constituída por um ou vários elementos colados e destinada a vedar os recipientes ou a contribuir para a sua estanquicidade (7.8 – TECH)
3	natural cork stopper Stopper consisting entirely of natural cork Note: Natural cork stoppers that have been submitted to the sealing operation (see 6.5.5) are commonly referred to as colmated natural stoppers. (5.5 – NORM)	rolha de cortiça natural Rolha totalmente constituída por cortiça natural. Nota: As rolhas naturais que tenham sido submetidas à operação de colmatagem (ver 6.5.5) são comummente designadas por rolhas naturais colmatadas. (5.5 – NORM)
4	colmated natural cork stopper The colmated natural cork stopper is a stopper made of natural cork in which its lenticels are filled with a mixture of glues and cork powder from the dimensional finishing processes of natural cork stoppers. (6.1 – REP)	rolha de cortiça natural colmatada A rolha de cortiça natural colmatada é uma rolha feita de cortiça natural em que são obturadas as suas lenticelas com uma mistura de colas e pó de cortiça proveniente dos acabamentos dimensionais das rolhas de cortiça natural. (6.1 – REP)
5	agglomerated cork stopper Stopper obtained by the agglutination of cork granules with a size between 0,25 mm and 8 mm, with addition of binders, by means of extrusion or moulding and composed of at least 51% by weight of cork granules. (5.5 – NORM)	rolha de cortiça aglomerada Rolha obtida pela aglutinação de granulado de cortiça com dimensão compreendida entre 0,25mm e 8mm, com adição de ligantes, através de extrusão ou moldagem e composta, pelo menos, por 51 % de granulado de cortiça, em peso. (5.5 – NORM)
6	agglomerated stopper: piece of agglomerated cork, obtained by extrusion or moulding (3.1 – STUD)	rolha aglomerada: peça de cortiça aglomerada, obtida por extrusão ou moldagem (3.1 – STUD)
7	n+n stopper Stopper formed by a body of agglomerated cork and "n" disks of natural cork glued to one or both ends. N.B.: In this designation, "n" indicates the number of disks used. (5.5 – NORM)	rolha n+n Rolha formada por um corpo de cortiça aglomerada e "n" discos de cortiça natural colados num ou em ambos os topos. Nota: Nesta designação, "n" indica o número de discos utilizados. (5.5 – NORM)
8	technical stopper Technical stoppers are composed of a very dense body of agglomerated cork with disks of natural cork glued to one end - or to both ends. Technical stoppers with one disk on each end are called 1+1 technical stoppers; those with two disks of natural cork on each end are called 2+2 technical stopper;	rolha técnica As rolhas técnicas são constituídas por um corpo de cortiça aglomerada, muito denso, com discos de cortiça natural colados no seu topo – ou em ambos os topos. As rolhas técnicas com um disco em cada topo são designadas rolhas técnicas 1+1. Com dois discos de cortiça natural em cada topo chamam-se

	and those with two disks glued at only one of the ends are called 2+0 technical stoppers. (6.1 – REP)	rolhas técnicas 2+2, e com dois discos em apenas um dos topos chamam-se rolhas técnicas 2+0. (6.1 – REP)
9	rounded stopper Stopper whose edges of one or two ends were rounded by abrasion. (5.5 – NORM)	rolha boleada Rolha cujas arestas de um ou dois topos foram arredondadas, por abrasão. (5.5 – NORM)
10	marked stopper Stopper whose lateral surface or ends were marked in ink or by fire (7.6 – TECH)	ROLHA MARCADA Rolha cuja superfície lateral ou topos foram marcados a tinta ou a fogo. (7.6 – TECH)

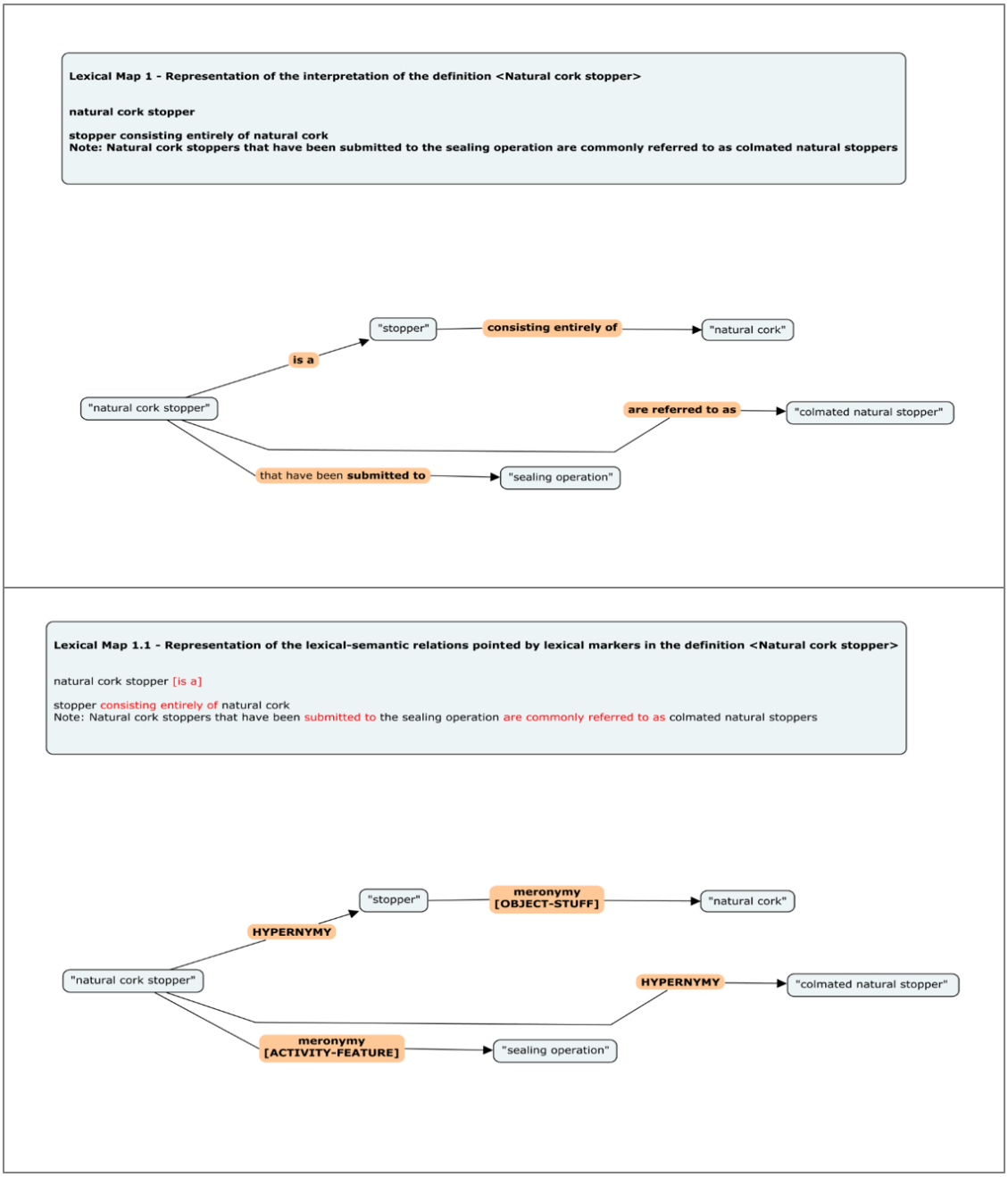
For this paper, we will consider only one definition, namely <Rolha de cortiça natural> [natural cork stopper] (see line 3 in Table 4), to demonstrate our linguistic and conceptual analysis. However, instead of using the definitional statement written in Portuguese, we have decided to use its literal translation into English for clarity.

Error! Reference source not found. **5**
Linguistic analysis of the definition of <Natural cork stopper>

Concept				
<Natural cork stopper>				
Definition in context				
stopper consisting entirely of natural cork Note: Natural cork stoppers that have been submitted to the sealing operation (see 6.5.5) are commonly referred to as colmated natural stoppers (Literal translation). Source: (Cork Corpus 5.5 – NORM)				
LINGUISTIC DIMENSION	**Analysis**	**Lexical marker (LM)**	**Lexical-semantic relations**	**Interpretation**
	natural cork stopper [is a] stopper	'is a' = Ø	**HYPERNYMY - HYPONYMY**	stopper [GENERIC] natural cork stopper [SPECIFIC]
	natural cork stopper [consists entirely of] natural cork	'consisting entirely of'	**HOLONYMY-MERONYMY**	natural cork stopper [OBJECT] natural cork [STUFF]
	natural cork stopper [is submitted to] the sealing operation	'submitted to'	**HOLONYMY-MERONYMY**	sealing operation [ACTIVITY] ? = [FEATURE]
	colmated natural stopper [is a] natural cork stopper	'commonly referred to as' same as = 'is a'	**HYPERNYMY - HYPONYMY**	natural cork stopper [GENERIC] colmated natural stopper [SPECIFIC]
	colmated natural stopper [results from] the sealing operation	results from = inferred from 'submitted to'	**HOLONYMY-MERONYMY**	sealing operation [ACTIVITY] colmated = [FEATURE]

Table 5 represents the first moment of our study, where we describe the deconstruction of the definition and present its linguistic analysis. The aim is to analyse the lexical-semantic relations between terms. The definition of <Natural cork stopper> is given in the main sentence, followed by some encyclopaedic information, namely the note. While the first sentence provides essential information for understanding what a <Natural cork stopper> is made of, the encyclopaedic information conveys information about what the object is when submitted to a specific operation. The first information that we obtain from the analysis is that a <Natural cork stopper> "is a stopper". In this statement, "is a" is a lexical marker that relates term A "natural cork stopper" and term B "stopper", giving us a clear hypernym-hyponym relation, where "natural cork stopper" is the hyponym of the hypernym "stopper".

In the second sentence – inserted as a note in the definition – another piece of information is obtained from the analysis of the statement "natural cork stoppers that have been submitted to sealing operation". Here, the lexical marker is "submitted to" [submetidas à] and relates the term "natural cork stopper" [rolha natural] to the term "sealing operation" [operação de colmatagem]. The term "sealing operation" – which indicates an operation/activity – is related by the lexical marker "submitted to" [submetidas à] to the term "natural cork stopper" – which we already know to be an object. The interpretation of their meanings allows us to infer that the lexical-semantic relation established is meronymy, subtype [ACTIVITY-FEATURE] [5] (see Map 1 for the former, and Map 1.1 for the latter, in Figure 3).



Error! Reference source not found.**3:** Lexical Map 1 and Lexical Map 1.1

4. The conceptual analysis

The conceptual analysis corresponds to the second stage of the analysis of the definition in focus. The differential characteristics found in this definition are expressed by /natural cork/, /natural/,

/colmated/ and /sealing operation/. The observations of this analysis are systematised in Table 6 and are based on the lexical markers found in the linguistic analysis of the definition. At the same time, based on the linguistic interpretation of the data, we extrapolated to conceptual relation identifiers.

Error! Reference source not found. **6**
The conceptual analysis of the definition of <Natural cork stopper>

	Aristotelian formula (X=Y+DC) X [species] = Y [genus] + DC [differential characteristic]					
CONCEPTUAL DIMENSION	**Analysis**	**Conceptual relation identifier**	**Conceptual relation**	**Interpretation**	**Transcription in X=Y+DC**	**Differential characteristics**
	natural cork stopper [is a] stopper	*is_a* [corresponds to LM *'is a'*]	**SUBSUMPTION**	stopper [GENUS] natural cork stopper [SPECIES]	natural cork stopper [SPECIES] = stopper [GENUS] + DC ?	
	natural cork stopper [is made of] natural cork	*has_substance* [corresponds to LM *'consisting entirely of'*]	**ASSOCIATIVE**	natural cork stopper [PRODUCT] natural cork [RAW MATERIAL]	natural cork stopper [SPECIES] = stopper [GENUS] + natural cork [DC]	*/natural cork/*
	natural cork stopper [is made of] natural cork	*has_substance* [corresponds to LM *'consisting entirely of'*]	**ASSOCIATIVE**	cork [MATTER] natural [PROPERTY]	natural cork [GENUS] = cork [GENUS] + natural [DC]	*/natural/*
	natural cork stopper [is submitted to] sealing operation	*has_process* [corresponds to LM *'submitted to'*]	**ASSOCIATIVE**	sealing operation = [PROCESS] ? = [RESULT]	? [SPECIES] = natural cork stopper [GENUS] + sealing operation [DC]	*/sealing operation/*
	colmated natural stopper [is a] natural cork stopper	*is_a* [corresponds to the LM *'commonly referred as'*]	**SUBSUMPTION**	natural cork stopper [GENUS] colmated natural stopper [SPECIES]	colmated natural stopper [SPECIES] = natural cork stopper [GENUS] + colmated [DC]	*/colmated/*

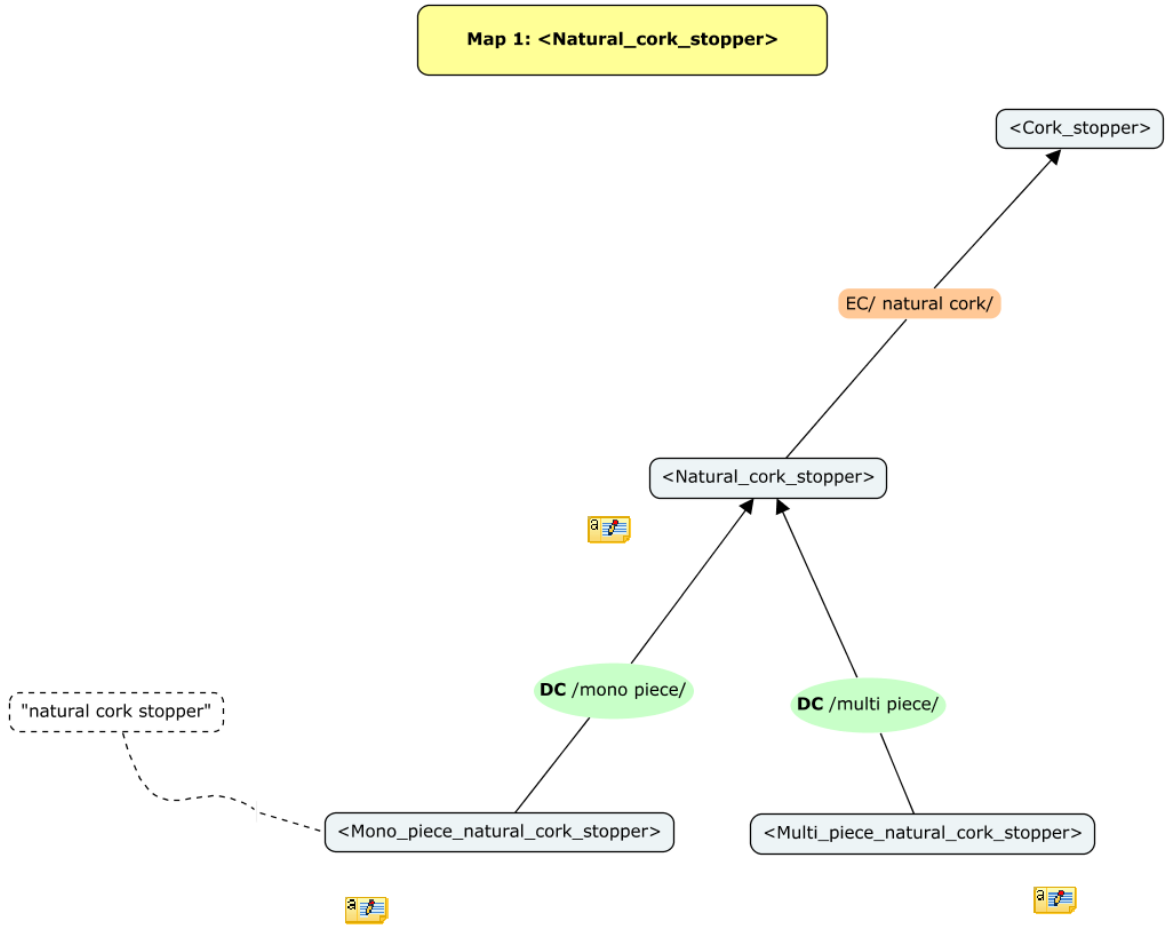
As systematised in Table 6, we propose three conceptual relation identifiers, namely, (1) has_substance, (2) is_a, and (3) has_process.

(1) *has_substance* is expressed by the lexical marker *"consisting entirely of"* [totalmente constituída por], which refers to the substance of the object. As we know from the linguistic analysis, the term "natural cork" points to the notion of substance, a material that a given object can be made of. Since <Stopper> is an object made of a substance, we propose the conceptual relation identifier *has_substance* to represent such a semantic relation. This semantic relation mirrors a pragmatic association - e.g., a thematic connection through virtue or experience, or a dependency between concepts established by the proximity of time and space [3] - in which a <Stopper> is a [PRODUCT] obtained from a substance, more specifically a [RAW MATERIAL]. From the interpretation of this information, we assume that an associative conceptual relation is in place, subtype PRODUCT – RAW MATERIAL, in which *stopper* points to the meaning of PRODUCT, and *natural cork* points to the meaning of RAW MATERIAL. This interpretation can be represented as follows: [stopper] $_{\text{PRODUCT}}$ *has_substance* [natural cork] $_{\text{RAW MATERIAL}}$.

The dichotomy PRODUCT – RAW MATERIAL has twofold importance at this point of the conceptual analysis: on the one hand, it underpins the subtype of the associative relation, while on the other hand, it is included in the Aristotelian formula [8],[6] known as X = Y + DC, where X=specific concept; Y=genus; and DC=differential characteristics. The purpose of using such a formula is to identify, for

the task of concept modelling, the characteristics stated in the definition under analysis. In order to use such a formula, one must first identify two concepts: the specific concept and its genus. [We will develop this further in the paper].

(2) *is_a* relation: <Natural Cork Stopper> is the subordinate concept, which we have labelled [SPECIES], and <Stopper> is the superordinate concept, which we have labelled [GENUS]. This assumption can be represented as: [natural cork stopper] SPECIES *is_a* [stopper] GENUS. Once the genus and the species have been identified, we can then insert these two elements in the formula X **SPECIES** = Y **GENUS** + DC, where: X = natural cork stopper; Y = stopper. Differential characteristics are inferred in a second stage: considering that [stopper] PRODUCT *has_substance* [natural cork] RAW MATERIAL, we can conclude that X [natural cork stopper] = Y [stopper] + DC [natural cork]. The first statement of the definition conveys the information represented by the first interpretation above, with the dichotomy [SPECIES-GENUS], which can be represented in the form of a conceptual map (see Figure 4). Conceptual map 1 is built by applying a *differentiae* dichotomy in which the differential characteristic /natural cork/ underlies one of the subdivision criteria[5].



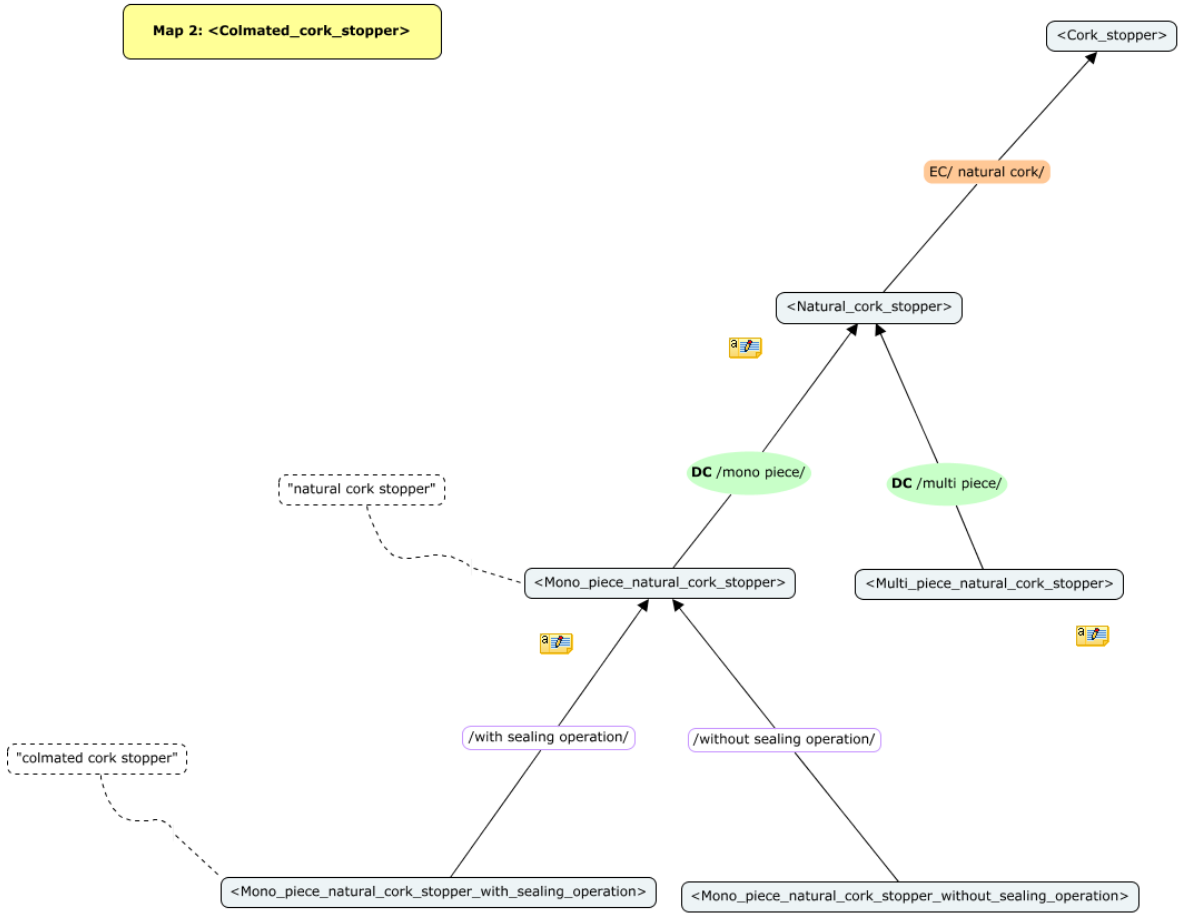
Error! Reference source not found.**4:** Conceptual Map 1 - two composition types of <Natural_cork_stopper> in CmapTools

Conceptual Map 1 (Figure 4) is the conceptual representation of the first statement of the definition, from which we have inferred that a <Natural cork stopper> *is_a* <Cork stopper>. Two axes of analysis are considered in this map: Substance and Parts (the 'Parts' axis was inferred from Definition 1; see Table 4). The conceptual information represented here, namely the axes of analysis Substance and Parts – whose underlying characteristics are /natural cork/, /mono piece/ and /multi piece/ – will be some of the coordinates for the elaboration of the formal description of the concept `NaturalCorkStopper` in

[5] According to (ISO/FDIS 1087), the "subdivision criterion [is the] type of characteristic according to which a superordinate concept is divided into subordinated concepts." (2019 (E), p. 5).

Protégé. Finally, <Multi_piece_natural_cork_stopper> will help us to formally describe types of <Stoppers> composed of several Parts – not only made of <Natural_cork>, but also of <Agglomerated_cork> and <Mixed_cork>. Here, the characteristics fall under the axis of analysis 'Parts' and are the coordinates for modelling multi-part concepts.

(3) *has_process*: Following the same method, the analysis of the note from which we obtained the information: <Natural cork stopper> is submitted to /sealing operation/, was represented in a second map (Figure 5). This piece of information grounds the conceptual relation identifier we have named as *has_process.*



Error! Reference source not found.**5:** Conceptual map of <Mono_piece_natural_cork_stopper_with_sealing_operation>

Conceptual Map 2 (Figure 5) is the representation of the two sentences of the definition in focus. Therefore, three axes of analysis are now considered: Substance, Parts, and Finishing Processes, to which the characteristics /with sealing operation/ and /without sealing operation/, were added. As represented in Conceptual Map 2, the characteristics /with sealing operation/ and /without sealing operation/ led us to a different level of concept representation, i.e., the concept <Mono_piece_natural_cork_stopper_with_sealing_operation>, verbally designated by "colmated cork stopper", is a specialisation of <Mono_piece_natural_cork_stopper>, in turn, verbally designated by "natural cork stopper". Therefore, these two concepts should not be treated at the same level, nor should they be defined in the same definitional context, either in natural language or in (semi)formal languages.

The conceptual relations we have inferred from the analysis of the lexical markers observed in the first five definitions (see Table 4), is summarised in Table 7.

Error! Reference source not found. **7**
Overview of the conceptual relations inferred from lexical markers

Lexical marker	Conceptual relation identifier	Conceptual relation	A typology of definitional texts governed by the DC
'is a'	*is_a*	SUBSUMPTION	stopper [SPECIES]= product [GENUS] + [any DC added to the genus]
'commonly referred as'	*is_a*	SUBSUMPTION	colmated natural stopper [SPECIES] = natural cork stopper [GENUS] + colmated [DC added to the genus]
'is a'	*is_a*	SUBSUMPTION	colmated natural cork stopper [SPECIES] = stopper [GENUS] + [any DC added to the genus]
'intended to'	*has_function*	ASSOCIATIVE	stopper [SPECIES] = product [GENUS] + to seal bottles [FUNCTION=DC]
'obtained from'	*has_raw_material*	ASSOCIATIVE	stopper [SPECIES] = product [GENUS] + natural cork [SUBSTANCE=DC]
'obtained from'	*has_raw_material*	ASSOCIATIVE	stopper [SPECIES] = product [GENUS] + agglomerated cork [SUBSTANCE=DC]
'obtained from'	*has_substance*	ASSOCIATIVE	natural cork [SPECIES] = cork [GENUS] + natural [SUBSTANCE=DC]
'obtained from'	*has_substance*	ASSOCIATIVE	natural cork [SPECIES] = cork [GENUS] + agglomerated [SUBSTANCE=DC]
'intended to'	*has_function*	ASSOCIATIVE	stopper [SPECIES] = piece of cork [GENUS] + to seal containers [FUNCTION=DC]
'piece of'	*has_substance*	ASSOCIATIVE	stopper [SPECIES] = piece [GENUS] + cork [SUBSTANCE=DC]
'usually'	*has_shape*	ASSOCIATIVE	stopper [SPECIES] = piece of cork [GENUS] + cylindrical [SHAPE=DC]
'usually'	*has_shape*	ASSOCIATIVE	stopper [SPECIES] = piece of cork [GENUS] + conical [SHAPE=DC]
'usually'	*has_shape*	ASSOCIATIVE	stopper [SPECIES] = piece of cork [GENUS] + prismatic quadrangular [SHAPE=DC]
'sometimes with'	*has_process*	ASSOCIATIVE	stopper [SPECIES] = piece of cork [GENUS] + rounded edges [PROCESS=DC]
'sometimes with'	*has_process*	ASSOCIATIVE	stopper [SPECIES] = piece of cork [GENUS] + chamfered edges [PROCESS=DC]
''consisting entirely of'	*has_substance*	ASSOCIATIVE	natural cork stopper [SPECIES] = stopper [GENUS] + natural cork [SUBSTANCE=DC]
'consisting entirely of'	*has_substance*	ASSOCIATIVE	natural cork [GENUS] = cork [GENUS] + natural [SUBSTANCE=DC]
'submitted to'	*has_process*	ASSOCIATIVE	? [SPECIES] = natural cork stopper [GENUS] + sealing operation [DC]
'is made of'	*has_raw_material*	ASSOCIATIVE	colmated natural cork stopper [SPECIES] = stopper [GENUS] + natural cork [SUBSTANCE=DC]
'is made of'	*has_substance*	ASSOCIATIVE	colmated natural cork stopper [SPECIES] = natural cork stopper [GENUS] + colmated [SUBSTANCE=DC]
'its lenticels are filled'	*has_process*	ASSOCIATIVE	colmated natural cork stopper [SPECIES] = natural cork stopper [GENUS] + filled lenticels [PROCESS=DC]
'results from'	*has_process*	ASSOCIATIVE	cork powder [SPECIES] = natural cork [GENUS] + dimensional finishing process [PROCESS=DC]
'consisting of'	*has_part*	PARTITIVE	stopper [SPECIES] = product [GENUS] + one piece [PARTS=DC]

'obtained from'	*has_part*	PARTITIVE	stopper [SPECIES] = product [GENUS] + several pieces [PARTS=DC]
'consisting of'	*has_part*	PARTITIVE	stopper [SPECIES] = piece of cork [GENUS] + one element [PARTS=DC]
'consisting of'	*has_part*	PARTITIVE	stopper [SPECIES] = piece of cork [GENUS] + several elements [PARTS=DC]

As shown in Table 7, differential characteristics (DC) can be any characteristic in a given definition according to the formula of an intensional definition [4], so that, depending on what is added to the intension of the [GENUS], the understanding of the concept's place in the concept system is provided. The same happens with the associative relation, although with several other axes of analysis involved. Here, DC share semantic labels in a more productive variety, namely [SUBSTANCE]; [FUNCTION]; [PROCESS] and [SHAPE], given the prolific semantic relations identified between concepts.

5. Building the ontology

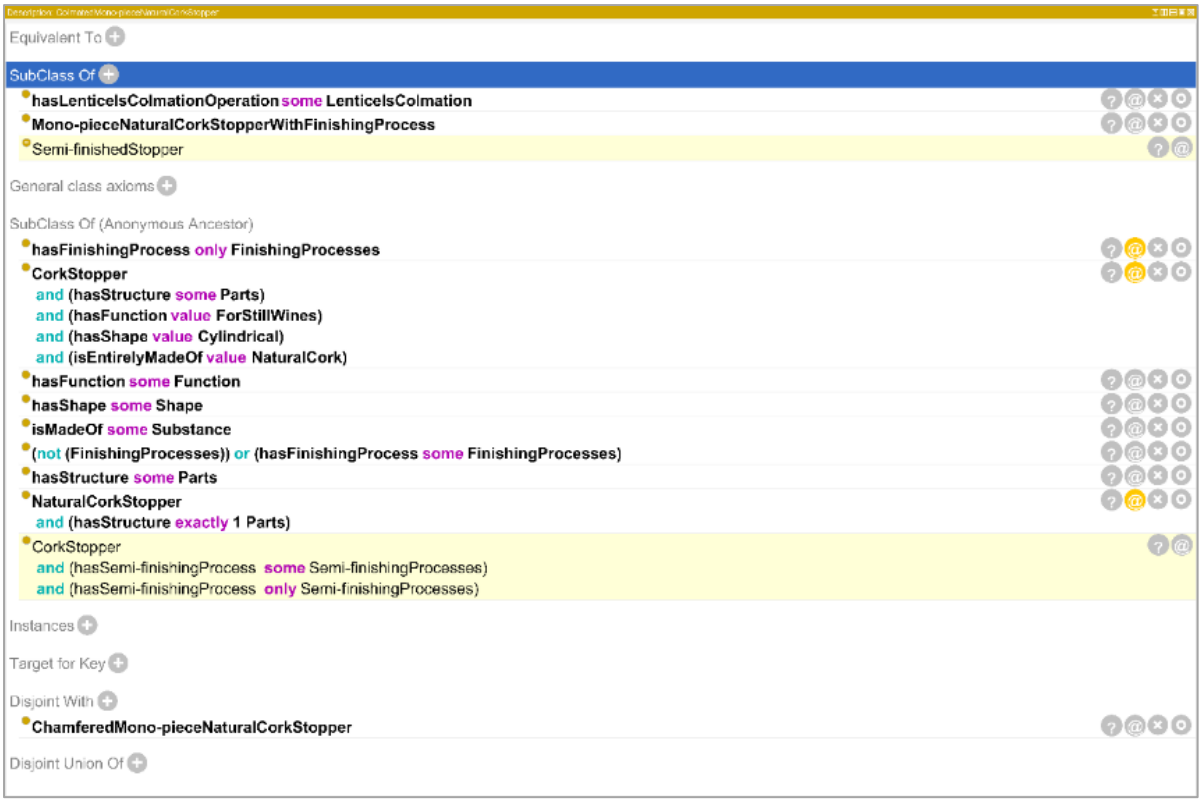
For the task of building OntoCork, we used the editor Protégé [7]. OntoCork is an ontology in which the concepts of the domain of cork are systematised through logical constructs. The descriptive domain properties (conceptual relations) elaborated to develop the ontology ground on the five axes of analysis that we have previously retained, namely Part, Substance, Shape, Finishing Process, and Function, as systematised in Table 8.

Error! Reference source not found. **8**
Five core conceptual relations of the ontology

Axis of analysis	Format in Protégé	Type of conceptual relation
FUNCTION	hasFunction	associative relation, subtype [OBJECT-FUNCTION]
SUBSTANCE	IsMadeOf	associative relation, covering both subtypes [RAW MATERIAL – PRODUCT] and [MATTER/SUBSTANCE – PROPERTY]
PARTS	hasStructure	partitive relation [PART-WHOLE]
FINISHING PROCESS	hasProcess	associative relation, within the subtype [PROCESS-RESULT]
SHAPE	hasShape	associative relation, subtype [OBJECT-SHAPE]

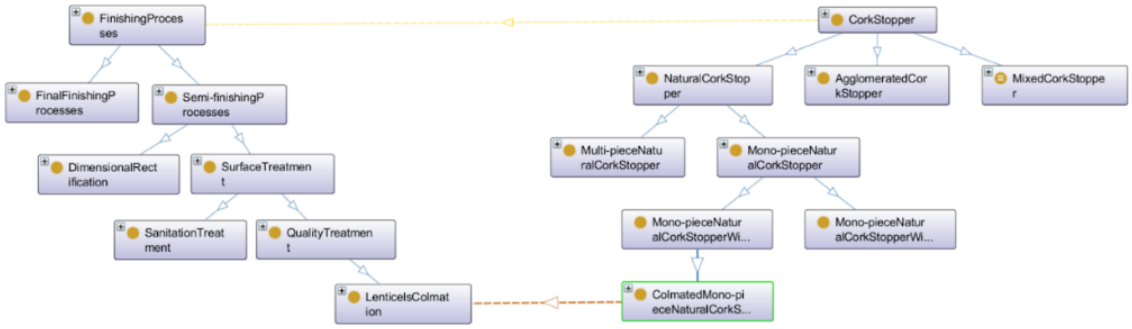
For this paper, we will present the description of the characteristics that build up the formal definition of <Natural cork stopper>, a closure with a body-structure of 1 <Part> submitted to <Sealing process>, in addition to the classification provided by the reasoner HermiT [6] as a <Semi-finished> object (see Figure 6).

[6] HermiT – a plugin reasoner of Protégé (http://www.hermit-reasoner.com/)



Error! Reference source not found.**6:** Concept description of ColmatedMonoPieceNaturalCorkStopper, in Protégé

Figure 7 is the ontological representation of `ColmatedMonoPieceNaturalCorkStopper`, in Ontograf[7], where we can observe several concepts systematised, either vertically: in a hierarchical dependency, or horizontally: in a pragmatic (associative) dependency, according to the differential characteristics. For clarity, we have decided to elide the visualisation of the associative relations between concepts that are not in focus in the following lines.



Error! Reference source not found.**7:** Ontological representation of ColmatedMonoPieceNaturalCorkStopper, in Ontograf

As illustrated in Figure 7, the `ColmatedMonoPieceNaturalCorkStopper` is a specification of `MonoPieceNaturalCorkStopperWithFinishingProcess`. The subsumption relation is represented by vertical blue arcs, and the associative relations are represented by horizontal dashed lines. The concepts

[7] https://protegewiki.stanford.edu/wiki/OntoGraf

ColmatedMonoPieceNaturalCorkStopper and LenticelsColmation are linked by the associative relation, subtype [PROCESS-RESULT]: hasLenticelsColmationOperation. This conceptual relation is based on the differential characteristic /with sealing operation/, which was drawn from the analysis of the definition of <Natural cork stopper>. Thus, hasLenticelsColmationOperation is the associative relation that induces the specification of MonoPieceNaturalCorkStopperWithFinishingProcess by differentia. Finally, it is also possible to see a hierarchical representation of FinishingProcesses, in which the involved operation of the concept we have just described is assigned as the most specific concept of this hierarchy. The interpretation of this subsumption is: LenticelsColmation is a kind of QualityTreatment, which is a kind of SurfaceTreatment, which in turn is a kind of Semi-finishingProcess, all of these are kinds of FinishingProcesses.

6. Conclusion

With this research, we wanted to explain the method used to build an ontology from human analyses of linguistic data. Linguistic and conceptual levels of analysis are to be analysed in relation to one another but as distinct phenomena. Texts are vehicles for knowledge transfer. Analysing texts to extract the characteristics of concepts, linguistically expressed by lexical markers pointing to lexical-semantic relations, allowed us to effectively capture the conceptual relations that are specific to the domain through the formula X=Y+DC. As demonstrated in this study, we were able to propose a preliminary conceptual organisation of the subject field. We have bridged three main aspects in our study: (i) the classical aspects of the Aristotelian logic; (ii) the methodology of our terminological work – where characteristics play a fundamental role in the analysis or the drafting of intensional definitions; and (iii) the formal definitions, for which we have used Protégé and the inherent Web Ontology Language (OWL) [12] to formally describe the concepts of the domain in order to relate them via abstract syntaxes and thus achieve formal reasoning, as concepts are consistently defined in a 'reason-able' ontology.

In future work, we intend to model the conceptual and the linguistic information contained in the resources we have developed, namely the ontology, the corpus, and a glossary (in progress) developed with Lexonomy[8], as linked data with the use of interoperable Linked Open Vocabularies[9].

7. References

[1] Atkins, S., Clear, J., & Ostler, N. (1992). Corpus Design Criteria. *Literary and Linguistic Computing, 7*(1), 1 - 16. doi:10.1093/llc/7.1.1

[2] Baker, P., Hardie, A., & McEnery, T. (2006). *A Glossary of Corpus Linguistics.* Edinburgh: Edinburgh University Press.

[3] ISO 704. (2009). Travail terminologique - Principes et méthodes. *NF ISO 704, 1er tirage 2009-12-P*. La Plaine Saint-Denis: Association Française de Normalisation.

[4] ISO/FDIS 1087. (2019 (E)). Terminology work and terminology science - Vocabulary. Suisse: ISO.

[5] L'Homme, M. C. (2004). *La Terminologie: principes et techniques - Paramètres.* Montréal, Canadá: Les presses de l'Université de Montréal.

[6] Meyer, I. (2001). Extracting Knowledge-Rich contexts for terminography: a conceptual and methodological framework. In D. Bourigault, C. Jacquemin, & M.-C. L'Homme (Eds.), *Recent Advances in Computational Terminology* (Vol. 2, pp. 279 - 302). Amsterdam / Philadelphia: John Benjamins B.V.

[7] Musen, M. A. (2015). The Protégé project: A look back and a look forward. doi:10.1145/2557001.25757003

[8] Pearson, J. (1998). *Terms in context.* Amsterdam: John Benjamins B.V.

[9] Pottier, B. (1992). *Théorie et analyse en Linguitique* (2, corrigée ed.). Paris: HACHETTE, Supérieur.

[8] https://github.com/elexis-eu/lexonomy
[9] https://lov.linkeddata.es/dataset/lov/

[10]Ramos, M. (2020). *Knowledge Organization and Terminology: application to Cork.* Lisboa: NOVA FCHS; LISTIC - Université Savoie Mont Blanc. Obtido de http://hdl.handle.net/10362/111722
[11]Ramos, M. (2020). OntoCork. NOVA FCSH. doi:https://doi.org/10.34619/a27q-1ryd
[12]W3C. (2004). *OWL Web Ontology Language Reference.* (M. Dean, & G. Schreiber, Eds.) Retrieved from W3C Recommendation 10 February 2004: https://www.w3.org/TR/owl-ref/