Extração de contextos definitórios do *Corpus COVID-19* com CQL

Definition context extraction from the COVID-19 corpus with CQL

Ana Eliza Pereira Bocorny[*]
Rozane Rebechi[**]
Cristiane Krause Kilian[***]

Resumo: Termos representam os conceitos de um domínio e sua compreensão permite o acesso aos saberes contidos nos textos especializados. Entender o significado dos termos, portanto, é de grande importância não apenas para que pesquisadores possam socializar seus estudos e descobertas, mas também para que profissionais e estudantes de várias áreas possam se valer da informação especializada em contextos de estudo e de trabalho. A evolução rápida do conhecimento muitas vezes não permite que a terminologia criada para designar conceitos seja dicionarizada com a necessária rapidez. Tal fato pode representar um grande desafio para aqueles que necessitam ter acesso ao conhecimento especializado. Tendo em vista o contexto descrito, este estudo parte da revisão de abordagens utilizadas para a extração automática de traços definitórios (TDs) e contextos definitórios (CDs) e propõe a utilização da ferramenta *Corpus Query Language* (*CQL*) para a extração de informações que auxiliem no entendimento da terminologia empregada em textos especializados. Em especial, verificamos a utilidade das sintaxes de busca construídas com a *CQL* para esse propósito, aplicando-as ao *Corpus COVID-19*. O percurso apresentado neste estudo poderá auxiliar não apenas especialistas da área médica, mas também tradutores, lexicógrafos e professores a processarem, de forma mais rápida e precisa, o conhecimento contido em textos especializados.

Palavras-chave: COVID-19; terminologia; Linguística de *Corpus*; extração automática de contextos definitórios (CDs); traços definitórios (TDs).

[*] Professora no Instituto de Letras da Universidade Federal do Rio Grande do Sul. E-mail: ana.bocorny@gmail.com.
[**] Professora no Instituto de Letras da Universidade Federal do Rio Grande do Sul. E-mail: rozane.rebechi@ufrgs.br.
[***] Professora no Instituto Superior de Educação Ivoti. E-mail: cristianekkilian@gmail.com.

Abstract: Terms represent the concepts of a domain and by comprehending them readers have access to the knowledge contained in specialized texts. Therefore, understanding the meaning of terms is of great importance not only for researchers to share the results of their studies, but also for professionals and students from various areas to apply specialized information in their learning and working contexts. The fast-evolving knowledge does not always permit that the terminology created to designate new concepts is quickly inserted in dictionaries, and this may represent a great challenge for those who need access to specialized knowledge. After presenting approaches used in the last twenty years for the automatic extraction of definition traits (DT) and definition contexts (DC), we propose the use of the Corpus Query Language (CQL) tool to retrieve information that helps in understanding the terminology used in specialized texts. In particular, we attested the usefulness of search syntaxes built with CQL for this purpose, applying them to the COVID-19 Corpus. The path presented in this study can help not only specialists in the medical field, but also translators, lexicographers and teachers to process, in a faster and more accurate way, the knowledge contained in specialized texts.

Keywords: COVID-19; terminology; *corpus* linguistics; definition context (DC) extraction; definitional segments (DS).

Introdução

Termos, ou unidades terminológicas, entendidos aqui como unidades lexicais com sentido especializado, são vetores de conhecimento. Eles representam os conceitos de um domínio e sua compreensão permite o acesso aos saberes contidos nos textos especializados. Ter domínio da terminologia correta é de grande importância para os pesquisadores que compartilham seus estudos e descobertas, e também para os profissionais de diversas áreas que se valem da informação especializada em seus contextos de trabalho. Novos conceitos, e, consequentemente, novos termos, são cunhados constantemente e os materiais terminográficos não são atualizados com a necessária rapidez, o que pode dificultar o acesso ao conhecimento especializado.

A pandemia do novo coronavírus é um exemplo da situação descrita. Pressionados pela evolução rápida da doença, pela publicação de um grande volume de pesquisas e pela necessidade de encontrar tratamento, prevenção e cura, pesquisadores e profissionais envolvidos com a pandemia deparam-se com necessidades terminológicas específicas que precisam ser atendidas de forma

rápida e precisa. Tais necessidades, via de regra, estão relacionadas à compreensão do significado de termos e unidades terminológicas. Além dos pesquisadores e profissionais mencionados, outros sujeitos, como tradutores, lexicógrafos, professores universitários e alunos de graduação e pós-graduação, também apresentam necessidades terminológicas específicas que surgem em função da pandemia.

As necessidades terminológicas variam, pois estão atreladas a vários fatores, como, por exemplo, ao perfil de cada sujeito, aos seus conhecimentos prévios e aos seus contextos de atuação. Um pesquisador com muito conhecimento especializado na área e um alto nível de proficiência na língua inglesa, por exemplo, possivelmente terá necessidades lexicográficas diferentes de um tradutor que atue em uma agência governamental e que não seja especializado na área em questão, mas que tenha um alto nível de proficiência em língua inglesa. Da mesma forma, um professor universitário da área de Letras que trabalhe leitura e escrita acadêmicas com alunos de graduação do curso de Medicina em uma instituição de ensino superior brasileira, tendo, supostamente, um baixo nível de conhecimento especializado e um alto nível de proficiência em língua inglesa geral e para fins acadêmicos, terá necessidades lexicográficas diferentes das de seus alunos, que, por outro lado, estão em fase de aquisição de conhecimento especializado da área médica e possuem níveis variados de proficiência em língua inglesa.

Como mencionado, todos esses sujeitos têm necessidades terminológicas diversas e processam a informação obtida de maneira diferente em função de saberes prévios e experiências variadas que constituem seu perfil. Todos eles, no entanto, poderiam se beneficiar de um recurso que, de forma intuitiva, ágil, rápida e precisa pudesse apresentar uma compilação de fragmentos de textos autênticos que contivessem informações sobre características, funções e aspectos relevantes para a construção e entendimento do significado dos termos relacionados a suas necessidades (*cf*. SIERRA 2009).

Após a discussão de algumas abordagens utilizadas para a extração automática de traços e contextos definitórios, visamos, neste estudo, apresentar estratégias de busca de informações que auxilie especialistas,

profissionais e estudantes no entendimento do significado de termos e unidades terminológicas de forma rápida e precisa. Em especial, buscou-se verificar a utilidade das sintaxes de busca (SB) construídas com a ferramenta *Corpus Query Language* (*CQL*) (KILGARRIFF *et al*. 2004) para esse propósito, a partir de um *corpus* especializado em língua inglesa. Para tanto, a partir da revisão dos conceitos de definição (D), traço definitório (TD) e contexto definitório (CD), e do levantamento de diferentes abordagens para a extração automática de definições, utilizou-se o *Corpus COVID-19* para testar a utilidade das sintaxes mencionadas. Esse *corpus* de artigos científicos escritos em inglês sobre a temática COVID-19, com aproximadamente 156.756.091 de palavras, é parte do *COVID-19 Open Research Dataset (CORD-19)*, e está disponível na ferramenta *Sketch Engine* (KILGARRIFF *et al*. 2004).

O estudo proposto nos levou a três questões de pesquisa: (i) Quais são os padrões definitórios mais produtivos do *corpus* de estudo? (ii) É possível construir SBs a partir dos padrões definitórios mais produtivos de uma área de especialidade? (iii) As SBs construídas com a *CQL* da ferramenta *Sketch Engine* (SE) são capazes de extrair TDs e CDs de *corpora* especializados, de forma a contribuir para o entendimento de termos e de unidades terminológicas?

Após esta introdução, o restante do artigo está organizado em quatro seções principais. A seção 1 inicia com a revisão da literatura relativa aos conceitos de D, CD e TD e às estruturas discursivas dos CDs. Na seção 2 são apresentadas algumas abordagens utilizadas para a extração automática de TDs e CDs. Em seguida, na seção 3, descrevemos os *corpora* utilizados na pesquisa e as etapas metodológicas seguidas. Na seção 4, apresentamos os resultados e a discussão referentes ao uso de SBs construídas a partir de padrões definitórios produtivos do *corpus* de estudo para extração de TDs e CDs do *Corpus COVID-19*. Por fim, trazemos as considerações finais e propostas de trabalhos futuros.

1. Conceitos de definição, contexto definitório e traço definitório

Nesta seção, trazemos algumas considerações sobre os conceitos de definição (D), contexto definitório (CD) e traço definitório (TD) e refletimos sobre as formas como os TDs e CDs podem auxiliar na compreensão da terminologia de uma área de especialidade. Em seguida, tratamos da identificação e descrição de algumas das estruturas discursivas possíveis dos CDs.

A elaboração de definições, uma das etapas mais complexas do processo lexicográfico, envolve muitas decisões metodológicas. Há, basicamente, três processos possíveis para a construção de definições. Elas podem ser: (i) criadas por especialistas; (ii) copiadas de outras obras de referência; ou (iii) extraídas de textos autênticos, como artigos acadêmicos, teses, dissertações, normas, leis, textos jornalísticos etc.

Em um entendimento amplo, pode-se dizer que a definição é o conjunto de informações dadas sobre o termo ou palavra. Trata-se, portanto, da explicação de um item lexical por meio de outros, possibilitando o entendimento do significado do item definido. Segundo Finatto (1998), as definições podem ser lexicográficas, enciclopédicas ou terminológicas. Suas características são guiadas tanto pelo tipo de dicionário do qual fazem parte - dicionário de língua geral, dicionário enciclopédico e dicionário especializado – quanto pela "instância" a qual se referem, seja uma palavra, um referente ou uma porção de conhecimento sobre coisas ou fenômenos.

Na definição lexicográfica, há predominância de informações linguísticas. Nos dicionários de língua geral, supostamente são elencadas todas as acepções de uma palavra, seu uso mais geral, mas também mais específico, regional e, por vezes, como parte de expressões idiomáticas. A definição enciclopédica "se ocupa mais de referentes e de descrição de 'coisas'" (FINATTO 1998: 135), apresentando informações de diversos tipos. Já a definição terminológica privilegia uma das acepções, apresenta o significado específico de um termo como parte de um sistema conceitual, de uma área específica,

delimitando-o e distinguindo-o das outras noções do mesmo sistema conceitual. Segundo Finatto (1998), esse tipo de definição "surge da complexa combinação de uma série de fatores, tais como as necessidades de veiculação de determinada porção de conhecimento e o perfil epistemológico e textual da área de especialidade" (FINATTO 1998: 138).

Vale ressaltar que a definição apresenta diferenças dependendo do tipo de obra na qual será incluída e de seus propósitos. Como aponta Barros (2004), "não existe uma definição válida para dois dicionários, uma vez que a cada tipo de obra correspondem algumas características específicas que determinam o conteúdo e a organização do enunciado definicional" (BARROS 2004: 159).

Quanto ao uso de definições em textos, Pearson (1996) ressalta que, quando especialistas escrevem artigos acadêmicos, eles podem definir (usando "definições originais"[1]) ou redefinir (usando "definições relatadas"[2]) termos. A autora afirma ainda que,

> se [os especialistas] criam e nomeiam um novo conceito, provavelmente definirão o termo quando o apresentarem pela primeira vez. Se um conceito e um termo já existem dentro de um domínio específico, um autor pode desejar expandir ou redefinir o conceito subjacente ao termo, alterando assim a definição.[3] (PEARSON 1996: 818).

Alguns autores fazem a distinção entre D e CD. Geralmente, entende-se contexto como porção de texto extraída de textos autênticos. Segundo De Besse (1991), contexto corresponde ao entorno linguístico de um termo e é constituído pelo enunciado ao seu redor. O CD possui duas funções: esclarecer o significado de um termo e exemplificar seu uso. O autor menciona vários tipos de contextos, distribuídos em dois grupos: os contextos que se referem ao conceito e os contextos que se referem ao termo, ou seja, à forma. O CD se encontra no primeiro grupo e "é formado por um certo número de elementos

[1] No original: *original definitions*. Todas as traduções são de nossa autoria.
[2] No original: *reported definitions*.
[3] No original: "*Authors will define the terms they use for a number of different reasons. If they have created and named a new concept, they are likely to define the term when they first introduce it. If a concept and a term already exist within a particular subject domain, an author may wish to expand or redefine the concept underlying the term, thereby altering the definition*".

úteis e necessários para a descrição do conceito, mas insuficientes para a redação de uma definição"[4] (DE BESSE 1991: 112).

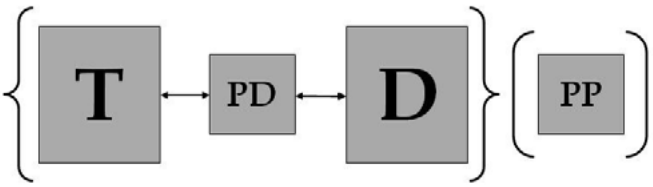
De Besse (1991) considera CD o que Pavel e Nolet (2002) denominam contexto explicativo: "Os CDs apresentam características essenciais do conceito em estudo, enquanto que os contextos explicativos fornecem informação sobre algumas das características" (PAVEL & NOLET 2002: 48). Em TERMIUM Plus®[5], banco de dados linguísticos e terminológicos do Governo Canadense, o contexto explicativo esclarece um ou mais de um aspecto da unidade terminológica, mas não contém elementos suficientes para constituir uma definição. Essa distinção também é feita em De Lucca (2006). Para o autor, o CD "apresenta descritores essenciais do conceito" (DE LUCCA 2006: 8).

Pearson (1998) menciona que, ao produzirem textos em contextos especializados, os autores possivelmente expliquem o significado de alguns termos usados. A quantidade de informação que constituirá essas explicações dependerá da disparidade existente entre o conhecimento do autor e do (suposto) leitor. Segundo a autora, as explicações podem ser fragmentos de informações presentes nos textos e necessitam ser recuperadas e organizadas para a formulação de definições dos termos.

Segundo Sierra (2009), o CD é uma estrutura discursiva formada por, no mínimo, dois elementos, o termo (T) e uma definição (D), os quais são conectados por um padrão definitório (PD)[6], que pode ser tipográfico (por exemplo, dois pontos, parênteses) ou sintático (por exemplo, 'se define como', 'é um'). Há também o elemento optativo, chamado de padrão pragmático (PPR), que apresenta informação metalinguística ou pragmática referente a forma, condições de uso ou alcance operativo (SIERRA 2009). Reproduzimos abaixo a figura que corresponde a essa estrutura:

[4] No original: "*contient un certain nombre d'éléments utiles et nécessaires à la description du concept, mais insuffisants pour la rédaction d'une definition*".
[5] http://www.btb.termiumplus.gc.ca/
[6] Também chamado de marcadores de definição (p. ex. MACIEL & FERREIRA 2005).

Figura 1: Estrutura de um CD



Fonte: Sierra (2009: 17).

Em nosso estudo, a partir do entendimento de Pearson (1998) e adaptando a proposta de Sierra (2009), chamamos de contexto definitório (CD) fragmentos de texto autêntico, contendo traços definitórios (TDs) e escritos por especialistas, ou seja, informações sobre características, funções e aspectos relevantes que auxiliam o leitor na construção do entendimento do significado dos termos. Diferentemente da definição, o CD não é um texto especialmente construído por lexicógrafos ou especialistas para compor uma obra lexicográfica ou terminográfica. Consideramos traço definitório (TD) a informação extraída de um contexto definitório que pode auxiliar na construção de uma definição ou na construção do significado do termo. Fornecem dados sobre características do termo, como composição, estrutura, função ou relação com outros termos.

Como exemplo, apresentamos o seguinte contexto definitório: *Novel coronavirus is a respiratory disease caused by a viral infection*. Ele contém dois traços definitórios – (i) *a respiratory disease* (doença respiratória) e (ii) *caused by a viral infection* (causada por uma infecção viral). O primeiro indica um hiperônimo e o segundo expressa uma causa.

Os desafios científicos e técnicos para a extração de CDs a partir de um *corpus* especializado são muitos. É necessário, em primeiro lugar, descrever as diferentes estruturas discursivas dos CDs para, só então, tratar das estratégias para sua extração. A metodologia proposta por Triki (2019) mostra a importância de se ter dados linguísticos precisos antes de iniciar o processo de extração de CDs. A autora parte de um processo manual de extração de segmentos definitórios (*definitional segments*), equivalentes ao que chamamos neste estudo de traços definitórios (TDs), realizado por especialistas a partir de um *corpus* de 40 artigos (317.000 palavras) de duas áreas de especialidade: Ciência da Computação e Linguística. O processo de codificação, descrito por

Triki (2019), iniciou com a leitura detalhada dos textos, durante a qual todas as definições foram marcadas. Em seguida, segmentos definitórios foram categorizados quanto a sua estrutura. Por fim, foi feita uma classificação do aspecto funcional das unidades definitórias (*defining units*) conforme as categorias estabelecidas por Triki (2014): nomeação, classificação, composição, função ou explicação.

Com base em Sierra (2009), classificamos os padrões definitórios em dois grupos: (i) padrões definitórios formais, que se referem a algum elemento formal - por exemplo, uso de dois pontos (:), de fonte diferente ou marcação em itálico ou negrito diferente para o termo e definição ou contexto definitório; e (ii) padrões definitórios linguísticos, relacionados ao uso de estruturas com verbos como *define, know* ou *call* (por exemplo, *x is defined as*) ou uso de marcadores reformulativos (por exemplo, *that is*). Para nossa análise, utilizamos os padrões definitórios sugeridos por Kovář, Močiariková e Rychlý (2016), como será demonstrado na próxima seção.

2. Abordagens para a extração automática de CDs

Como afirmam Kosem *et al.* (2019), a extração de exemplos para a construção de produtos lexicográficos pode ser baseada em três tipos de *input*: humano, por máquina ou ainda uma combinação de ambos. O mesmo acontece com a extração de CDs. Nesta seção, revisamos a literatura quanto às abordagens para a extração automática desses contextos. Não é nossa intenção fazer uma revisão exaustiva de todas as abordagens existentes, mas sim um levantamento dos principais estudos conduzidos entre 2000 e 2020 que tiveram como objetivo principal o desenvolvimento de processos para a extração automática de CDs.

As pesquisas em extração automática de Ds e CDs empregam diferentes métodos de identificação de informações em um texto. Entre 2000 e 2020, alguns estudos (por exemplo, KLAVANS & MURESAN 2000; CUI, KAN & CHUA 2004 e

2005; FAHMI & BOUMA 2006; JIN *et al*. 2013; KOVÁŘ, MOČIARIKOVÁ & RYCHLÝ 2016; e VEYSEH *et al*. 2020) desenvolveram e aprimoraram esses processos. Veyseh *et al*. (2020) sugerem que tais abordagens sejam divididas em três categorias, conforme o tipo de extração: (i) abordagem baseada em regras; (ii) abordagem baseada em engenharia de atributos (*feature engineering*); e (iii) abordagem baseada em aprendizagem profunda (*deep learning*). O Quadro 1 elenca os estudos citados, listando o nome dos pesquisadores que conduziram o projeto, o ano de publicação do artigo que apresenta o estudo e a abordagem adotada para a extração de Ds e CDs do *corpus*.

Quadro 1: Abordagens de extração automática de Ds e CDs

Pesquisadores / Desenvolvedores	**Ano de publicação do artigo**	**Abordagem adotada para extração das Ds e CDs**
Klavans e Muresan	2000	Abordagem baseada em regras
Cui, Kan e Chua	2004 e 2005	Abordagem baseada em regras
Fahmi e Bouma	2006	Abordagem baseada em regras
Jin *et al*.	2013	Abordagem baseada em engenharia de atributos
Kovář, Močiariková e Rychlý	2016	Abordagem baseada em regras
Veyseh *et al*.	2020	Abordagem baseada em aprendizagem profunda

Fonte: Elaborado pelas autoras.

O *Definder*, desenvolvido por Klavans e Muresan (2000), foi um dos primeiros sistemas de extração automática de termos e CDs. Baseado em um sistema de regras, ele busca artigos completos e deles extrai termos e definições. Em um estudo posterior, Klavans e Muresan (2001) propõem uma avaliação quantitativa e qualitativa do recurso. Kovář, Močiariková e Rychlý

(2016) sugerem uma forma automática de extração de definições também baseada em regras. Utilizando a linguagem de consulta de *corpus* (*CQL)*, os autores construíram SBs que representam padrões definitórios recorrentes no *corpus* analisado para localizar e armazenar candidatos a definições no *corpus*. *Corpus Query Language* (*CQL*) é a linguagem utilizada para extrair informações de um *corpus*. Na composição do termo *CQL*, a palavra *query* significa questionamento, consulta ou busca. Os questionamentos realizados por meio da *CQL* utilizam uma série de atributos que, combinados, resultam em uma sintaxe de busca (por exemplo, [word="*coronavirus*"]) que extrairá do *corpus* todas as linhas de concordância que contêm a informação solicitada. O Quadro 2 mostra alguns exemplos desses padrões de forma simplificada.

Quadro 2: Forma simplificada dos padrões definitórios usados no SE para identificar candidatos à definição.

- **TERM** "is/are/means/was/were" "a/an", including:
  - "**TERM**" (in quotes)
  - **TERM** parenthesis "is/are/..." "a/an" (parenthesis expressed by commas, dashes or brackets)
  - **TERM** prepositional-phrase "is/are/..." "a/an"
  - optional "a/an" in selected cases
- "What" "is" **TERM**, with a definition in the following sentence
- **TERM** "refers" "to", plus variants with parentheses and prepositional phrases, as above
- **TERM** "is/are" "defined" "as", plus variants with parentheses and prepositional phrases, as above
- ... "is/are" "known/called/referred to" "as" **TERM**
- **TERM** "is/are" "used" "to" "describe/denote/mean/refer to", plus variants with parentheses
- **TERM** "is" "a" "term" "for/referring to", plus variants with parentheses
- **TERM** "is/are" "understood" "to", plus variants with parentheses
- **TERM** "consists" "of", plus variants with parentheses

Fonte: Kovář, Močiariková e Rychlý (2016: 391).

Kovář, Močiariková e Rychlý (2016) ressaltam que, em todos os padrões apresentados no Quadro 2, **TERM** é um sintagma nominal. Os autores também mencionam que os padrões, em *CQL*, são bastante complexos. Um exemplo de padrão definitório em *CQL* é mostrado na Figura 2.

Figura 2: Exemplo de padrão definitório em CQL para **TERM** "is/are" "understood" "to".

```
( (<s> | (<s>[tag="DT" & lc!="this|these|those"]) |
   ([tag!="IN|PP.*|POS" | lc="while|although"]
             [tag="DT" & lc!="this|these|those"]) |
   ([tag!="DT|IN|PP.*|POS|N.*|JJ.*|VVG|CD" | lc="while|although"])
  )
  ([tag="N.*|JJ|VVG|CD"]{0,3} !containing
   (meet [tag="N.*|JJ|VVG"] [tag="IN" & lc!="while|although"] -1 0)
  )
  1:[tag="N.*" & lemma!="reference|use|...|name|definition"]
  "\"|'"?
  "is|are"
  [tag="RB"]?
  "understood"
  "to"
  [tag="VV|VB"]
  [tag!="N.*"]{0,12}
  2:[tag="N.*"])
) within  <s/>
```

Fonte: Kovář, Močiariková e Rychlý (2016).

A partir da identificação de padrões definitórios recorrentes, um link chamado *Definitions* foi incluído no SE. Tal link filtra os resultados de qualquer concordância a partir dos parâmetros estabelecidos nas SBs construídas com base nos padrões definitórios recorrentes identificados pelos autores. A Figura 3 mostra os padrões definitórios sugeridos por Kovář, Močiariková e Rychlý (2016), bem como o número de vezes que o padrão definitório descrito aparece no *corpus* (*No. of hits*), o percentual de linhas de concordância (LCs), em uma amostra de 50 linhas, que contêm definições (*Prec. on sample*) e, por fim, uma

estimativa do número de definições contidas no *corpus* a partir do percentual identificado na amostra observada (*Estimated no. definitions*).

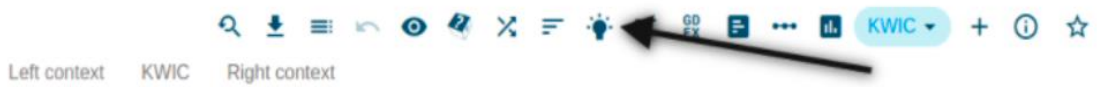
Figura 3: Padrões definitórios sugeridos por Kovář, Močiariková e Rychlý (2016).

Pattern type	No. hits	Prec. on sample (%)	Estimated no. definitions
is/are/...	1,751,813	77	1,342,215
what is	1,574	16	251
refers to	40,690	57	23,093
is defined as	54,435	29	15,781
is known as	78,756	54	42,528
is used to describe	3,934	63	2,485
is a term for	11,504	74	8,512
is understood to	618	35	216
consists of	12,821	29	3,721
Total	1,956,145	74	1,438,802

Fonte: Kovář, Močiariková e Rychlý (2016).

Até a publicação deste artigo, a funcionalidade *Definitions* estava disponível apenas para os *corpora English Wikipedia Corpus* e *enTenTen 13*. Nesses *corpora*, essa funcionalidade é identificada pelo ícone de uma lâmpada, como mostra a Figura 4:

Figura 4: Funcionalidade *Definitions* do SE.



Fonte: *Sketch Engine* (KILGARRIFF *et al*. 2004).

Tendo em vista o objetivo deste trabalho, que é oferecer aos especialistas estratégias de busca de informações que permitam o entendimento do significado de termos e unidades terminológicas, e considerando que muitos desses especialistas possivelmente não têm conhecimentos mais aprofundados de informática e de programação, é importante que a alternativa de busca proposta seja acessível e de fácil

implementação. Por essa razão, optou-se por testar a metodologia sugerida por Kovář, Močiariková e Rychlý (2016) de forma a verificar sua aplicabilidade para a extração de CDs e TDs de termos presentes em um *corpus* especializado.

3. O *corpus* e as etapas metodológicas

Nesta seção, a partir da aplicação de elementos da metodologia proposta por Kovář, Močiariková e Rychlý (2016) e de dados extraídos do *Corpus COVID-19*, exemplificamos análises realizadas para a identificação dos padrões definitórios mais produtivos no *corpus* de estudo, que pudessem ser usados para a extração de TDs e CDs. Para tanto, iniciamos descrevendo os *corpora*, e, em seguida, apresentamos os procedimentos metodológicos adotados.

3.1 Constituição do *corpus* de estudo

Os CDs e TDs foram coletados do *corpus* de estudo COVID-19. Tal *corpus* possui aproximadamente 156.756.091 de palavras e é composto de artigos científicos escritos em língua inglesa sobre a temática COVID-19. Tais artigos foram publicados em periódicos internacionais de livre acesso, após terem sido revisados por pares. Esse *corpus* é disponibilizado como parte do *COVID-19 Open Research Dataset (CORD-19)*[7]. O quadro abaixo apresenta a descrição do *corpus* de estudo:

[7] A fim de possibilitar pesquisas em processamento de linguagem natural (PLN) e inteligência artificial (IA) que contribuam com o combate à pandemia de COVID-19, a Casa Branca e importantes grupos de pesquisa construíram o *COVID-19 Open Research Dataset (CORD-19)*, um conjunto de mais de 51.000 artigos acadêmicos sobre COVID-19, SARS-CoV-2 e coronavírus, de livre acesso para os pesquisadores (MASUN et al. 2020).

Quadro 3: Descrição do *corpus* de estudo.

Registro[8]	Acadêmico
Gênero	Artigos
Meio de publicação	Periódicos revisados por pares e de acesso aberto da plataforma *CORD-19*
Língua de publicação	Inglês
Domínio	Medicina e ciências da saúde

Fonte: Elaborado pelas autoras.

O Quadro 4 mostra o número de *tokens*, *types* e textos do *corpus* de estudo:

Quadro 4: *Corpus* de estudo em números.

Domínio	***Corpus* de estudo**	**Número total de palavras (*tokens*[9])**	**Número total de palavras sem repetição (*types*)**	**Textos**
Medicina e áreas da saúde	Artigos científicos	224.061.570	1.783.529	50.754

Fonte: Elaborado pelas autoras.

[8] Biber, Connor e Upton (2007) definem *registro* (quando diferenciado de *gênero*) como a linguagem associada a uma área do conhecimento ou a um domínio, como o registro jurídico ou o registro acadêmico. O termo *gênero*, por sua vez, quando contrastado com *registro*, é usado para se referir a um tipo de mensagem com uma estrutura interna convencionalizada, como em um artigo de Biologia ou um memorando de negócios.

[9] A ferramenta SE denomina *token* a menor unidade existente em um *corpus*. Assim, diferentes formas das palavras e sinais de pontuação são contabilizadas como *tokens* distintos. Palavras unidas por apóstrofo e hífen são contabilizadas separadamente.

O *corpus* de estudo está disponibilizado no SE. A ferramenta foi utilizada para a extração das linhas de concordância contendo CDs e TDs de termos com alto índice de chavicidade no *Corpus COVID-19*. O *corpus* de referência usado foi o *English Web 2013 (enTenTen13)*, sugerido por *default* pelo SE. Esse *corpus* faz parte da família *TenTen corpus* e é composto de textos de língua geral coletados da internet.

O Quadro 5 mostra o número de *tokens*, *types* e textos do *corpus* de referência.

Quadro 5: *Corpus* de referência em números

Domínio	*Corpus* de estudo	Número total de palavras (*tokens*)	Número total de palavras sem repetição (*types*)	Textos
Língua geral	Textos extraídos da WEB	19.685.733.337	44.909.567	37.061.719

Fonte: Elaborado pelas autoras.

3.2 Etapas metodológicas

As etapas metodológicas adotadas neste estudo foram: (i) seleção de um *corpus* de estudo e um *corpus* de referência; (ii) identificação de termos e unidades terminológicas com maior índice de chavicidade e frequência normalizada (pmp)[10] no *corpus* de estudo; (iii) identificação de padrões definitórios apresentados na literatura (KOVÁŘ, MOČIARIKOVÁ & RYCHLÝ 2016); (iv) identificação dos padrões definitórios mais produtivos do *corpus* de estudo; (v)

[10] Neste caso, a frequência normalizada refere-se ao número de vezes que determinado item lexical aparece no *corpus* por milhão de palavras (pmp).

construção de sintaxes de busca, a partir dos padrões definitórios mais produtivos do *corpus* de estudo; (vi) extração de LCs com CDs e TDs de termos com alto índice de chavididade no *corpus* de estudo; e, por fim, (vii) observação e análise dos dados obtidos.

4. Resultados e discussão

Os procedimentos metodológicos para a extração dos CDs e TDs iniciaram com a identificação de termos do *Corpus COVID-19* com maior índice de chavicidade[11] e frequência normalizada (pmp). As palavras-chave identificadas são mostradas no Quadro 6:

Quadro 6: Termos simples e termos compostos com maior chavicidade e frequência normalizada (pmp) do *Corpus COVID-19*.

Termos simples	Índice de chavicidade	Frequência (pmp)
RNA	277.490	933,48
coronavirus	231.190	229,76
SARS	207.470	327,58
Termos compostos		
influenza virus	110.740	128,34
viral replication	93.530	97,95
respiratory tract	82.040	118,61

Fonte: Elaborado pelas autoras.

[11] Essa medida indica o quanto determinada unidade é mais frequente no *corpus* de estudo em comparação com o *corpus* de referência (BOCORNY *et al*. 2021).

Uma vez identificados os termos e as unidades terminológicas com maior chavicidade e frequência normalizada (pmp) do *corpus* de estudo, o termo *coronavirus* foi escolhido para exemplificar o tipo de análise proposto, a partir dos padrões definitórios sugeridos por Kovář, Močiariková e Rychlý (2016).

Diferentemente das metodologias descritas nas pesquisas citadas, neste estudo não buscamos o percentual de LCs com ocorrência de definições, mas sim a ocorrência de LCs que contivessem TDs e CDs. O Quadro 7 mostra os resultados extraídos do *Corpus COVID-19* para os padrões definitórios sugeridos por Kovář, Močiariková e Rychlý (2016) com o termo *coronavirus*:

Quadro 7: Padrões definitórios com o termo *coronavirus*.

Padrão definitório	Número total de ocorrências no *corpus*	Percentual de LC com TDs na amostra[12] (%)	Número estimado de LC com TDs no *corpus*
coronavirus *is a/an*	127 (0,45 pmp)	94%	119
what is ***coronavirus***	0	0%	0
coronavirus *refers/referred to*	10	100%	10
coronavirus *is defined as*	1	0%	0
coronavirus *is known as*	5 (0,02 pmp)	100%	5
coronavirus *is used to describe*	0	0%	0
coronavirus *is a term for*	0	0%	0
coronavirus *is understood to*	0	0%	0

[12] As 50 primeiras linhas de concordância extraídas com a *CQL* descrita constituem a amostra.

coronavirus *consists of*	5 (0,02 pmp)	100%	5

Fonte: Elaborado pelas autoras.

Para a identificação dos padrões definitórios mais produtivos do *corpus* de estudo, a partir dos padrões sugeridos por Kovář, Močiariková e Rychlý (2016), construímos as SBs usando a funcionalidade *CQL builder* disponível no *SE*. Exemplificamos a construção das sintaxes mostrando os atributos usados para o padrão definitório ***TERM*** *is/are a/an: [tag="N.*"] [word="is" | word="are"] [word="a" | word="an"]*. O atributo *tag* indica a classe gramatical da palavra que se quer extrair - no caso, *noun* (substantivo); e *word* indica a palavra exata que se busca - *is*. A barra vertical (|) também foi usada, indicando que se quer uma das opções que se encontram antes ou depois da marcação vertical.

Uma vez construídas todas as SBs, verificou-se a frequência absoluta e a frequência normalizada dos padrões, bem como o percentual de LCs com TDs em uma amostra de 50 LCs. Por fim, o número estimado de LCs com TDs foi calculado, como mostra o Quadro 8:

Quadro 8: Padrões definitórios extraídos do *Corpus COVID-19*.

Padrão definitório	Sintaxe de busca (SB)	Número total e pmp de LC extraídas do *corpus* com a *CQL* genérica	Percentual de LC com TDs na amostra (%)	Número estimado de LC com TDs no *corpus*
TERM *is*/are *a/an*	*[tag="N.*"] [word="is" | word="are"] [word="a" | word="an"]*	88.617 (315,6 pmp)	86%	74.490
what is ***TERM***	*[word="what"] [word="is" | word="are"] [tag="N.*"]*	44 (0,1 pmp)	5%	2
TERM	*[tag="N.*"] [lemma="refer"]*	3.342 (11,9	84%	2.807

refers/referred to	*[word="to"]*	pmp)		
***TERM** is defined as*	*[tag="N.*"] [word="is"] [word="defined"] [word="as"]*	3.003 (10,7 pmp)	100%	3.003
***TERM** is known as*	*[tag="N.*"] [word="is"] [word="known"] [word="as"]*	888 (3,1 pmp)	100%	888
***TERM** is used to describe*	*[tag="N.*"] [word="is"] [word="used"] [word="to"] [word="describe"]*	71 (0,2 pmp)	100%	71
***TERM** is a term for*	*[tag="N.*"] [word="is"] [word="a"] [word="term"] [word="for"]*	7 (0,02 pmp)	100%	7
***TERM** is understood to*	*[tag="N.*"] [word="is"] [word="understood"] [word="to"]*	24 (0,09 pmp	100%	24
***TERM** consists of*	*[tag="N.*"] [word="consists"] [word="of"]*	8.468 (30,1 pmp)	100%	8.468

Fonte: Elaborado pelas autoras.

De todos os padrões analisados, o que apresentou menor percentual (5%) de LCs com TDs na amostra analisada e uma das mais baixas frequências normalizadas (0,1 pmp) foi o padrão *what is **TERM***, destacado no Quadro 8 em azul. Outros padrões (***TERM** is known as, **TERM** is used to describe, **TERM** is a term for, **TERM** is understood to*), destacados no Quadro 8 em verde, ainda que tivessem um percentual alto (100%) de LCs com TDs, apresentavam frequências normalizadas baixas, menores que 10 ocorrências (pmp). Assim, para determinar os padrões definitórios mais produtivos do *corpus* de estudo, dois pontos de corte foram estabelecidos: (i) o padrão definitório deveria ter pelo menos 80% da amostra de LC com TDs e uma frequência normalizada

mínima de 10 ocorrências (pmp). Dessa forma, foram considerados produtivos no *Corpus COVID-19* os padrões definitórios ***TERM*** *is*/are *a*/*an*, ***TERM*** *refers*/*referred to*, ***TERM*** *is defined as* e ***TERM*** *consists of*, todos destacados em vermelho no Quadro 8.

Uma vez estabelecidos os padrões definitórios mais produtivos do *corpus* de estudo, procuramos verificar se haveria a possibilidade de agrupar vários padrões definitórios em uma única SB, de forma a extrair um maior volume de LCs com TDs. O agrupamento de padrões definitórios em uma mesma sintaxe de busca, inicialmente pretendido, aconteceu apenas com o padrão ***TERM*** *is defined as,* como pode ser observado no Quadro 9. Ao incluir na SB outros verbos, como *known, presented, treated, described, understood* e *used,* buscou-se extrair do *corpus* de estudo um volume maior de LCs com TDs. Por fim, percebeu-se a possibilidade de trocar todos os verbos listados pela etiqueta de verbos (*[tag="V.*"]*):

Quadro 9: Padrões definitórios mais produtivos do *corpus* de estudo.

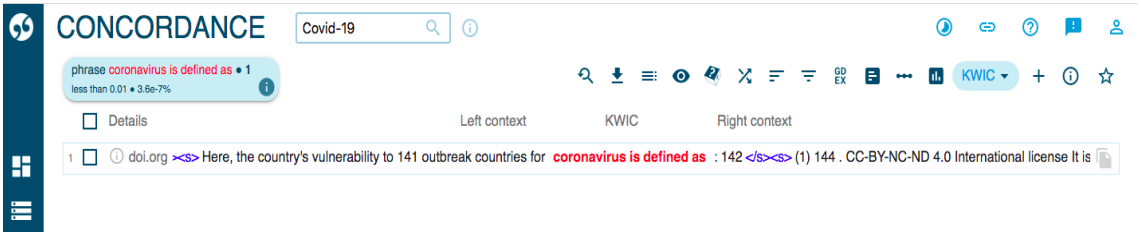
Padrão definitório	Sintaxe de busca
TERM *is*/are *a*/*an*	***TERM*** *[word="is" | word="are"] [word="a" | word="an"]*
TERM *refers*/*referred to*	***TERM*** *[lemma="refer"] [word="to"]*
TERM *is defined as*	***TERM*** *[word="is"] [word="are"] [tag="V.*"] [word="as"]*
TERM *is known as*	
TERM *consists of*	***TERM*** *[lemma="consist"] [word="of"]*

Fonte: Elaborado pelas autoras.

As Figuras 5 e 6 comparam, respectivamente, os resultados obtidos na extração de LC com TDs para o termo *coronavirus* com a combinação de palavras *is defined as* e com a SB *[word="coronavirus"] [word="is"] [word="defined" | word="presented" | word="known" | word="treated" |*

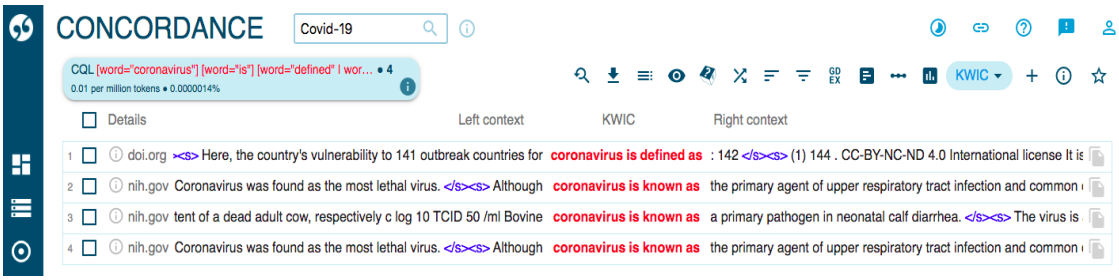
word="described" | *word="understood"* | *word="used"] [word="as"].* Ressalta-se que, com a combinação de palavras *is defined as*, apenas um CD foi extraído. A utilização da SB, por outro lado, permitiu a extração de quatro CDs:

Figura 5: Resultado da extração de CDs para o termo *coronavirus* com a combinação de palavras *is defined as*.



Fonte: *Sketch Engine* (KILGARRIFF *et al*. 2004).

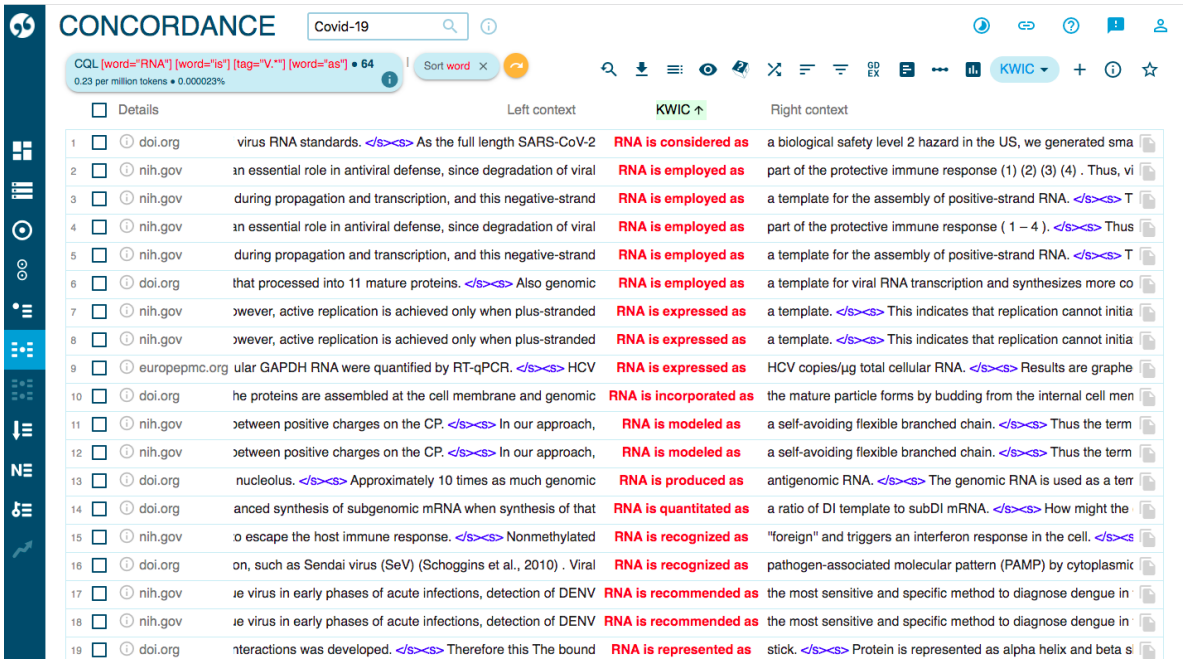
Figura 6: Resultado da extração de CDs para o termo *coronavirus* com a *SB [word="coronavirus"] [word="is"] [word="defined"* | *word="presented"* | *word="known"* | *word="treated"* | *word="described"* | *word="understood"* | *word="used"] [word="as"].*



Fonte: *Sketch Engine* (KILGARRIFF *et al*. 2004).

Na Figura 7, mostra-se a SB que substitui todos os verbos listados (*[word="defined"* | *word="presented"* | *word="known"* | *word="treated"* | *word="described"* | *word="understood"* | *word="used"])* pela etiqueta que representa verbos (*[tag="V.*"]*). Tal opção resulta não apenas em uma maior frequência normalizada (pmp), mas também em um maior percentual de LCs com CDs e TDs. Certamente, um volume maior de LCs com TDs e CDs facilitará o entendimento do significado do item lexical em questão.

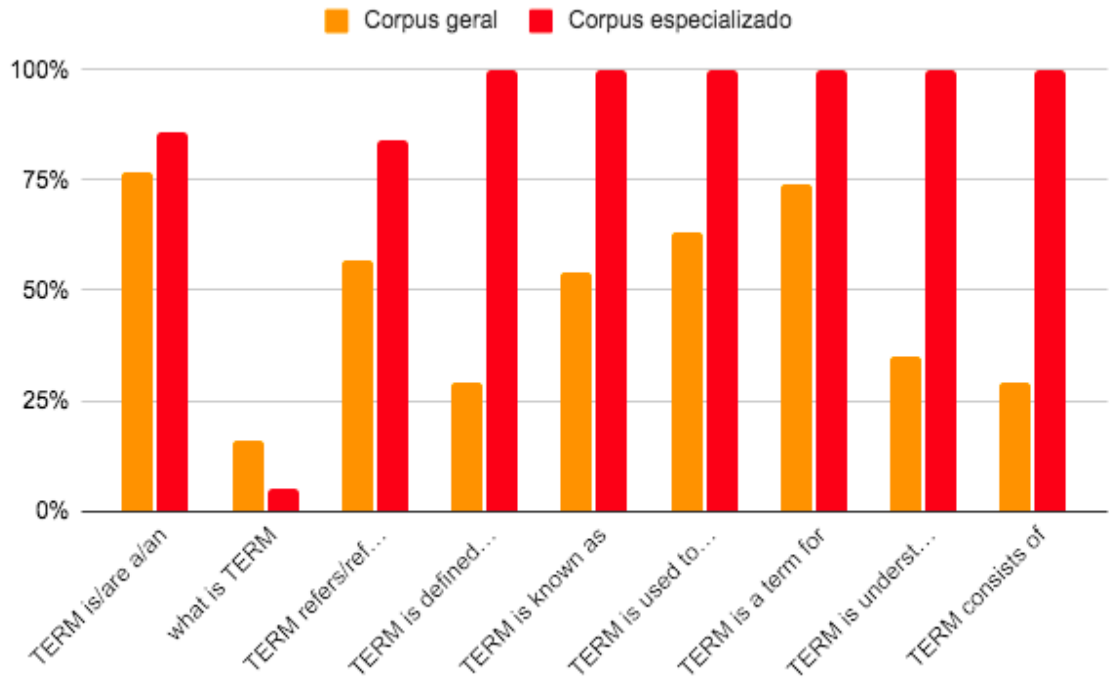
Figura 7: Resultado da extração de CDs do termo *coronavirus* com SB com *[tag="V.*"]*.



Fonte: *Sketch Engine* (KILGARRIFF *et al*. 2004).

Cabe destacar, também, que, com exceção de *what is **TERM***, os percentuais de LCs com CDs e TDs foram sempre maiores no *corpus* especializado, em comparação com o *corpus* geral, como pode ser observado no Gráfico 1. Tal fato não surpreende, pois é esperado que textos especializados, especialmente de temática recente, como a da COVID-19, tenham um volume maior de termos a serem definidos e ressignificados e, consequentemente, de contextos definitórios, do que textos da língua geral.

Gráfico 1: Percentuais de LC com TDs no *corpus* geral e no *corpus* especializado.



Fonte: Elaborado pelas autoras.

Considerações finais

O objetivo deste trabalho foi identificar estratégias de busca de informações que auxiliassem especialistas, profissionais e estudantes na compreensão do significado de termos e unidades terminológicas. Em especial, buscou-se verificar a utilidade das SBs construídas com a *Corpus Query Language* (*CQL*) para esse propósito, com base nos padrões definitórios sugeridos por Kovář, Močiariková e Rychlý (2016) e nas análises realizadas a partir do *Corpus COVID-19*.

Primeiramente identificamos os padrões definitórios mais produtivos do *corpus*: (i) ***TERM*** *[tag="N.*"] [word="is" | word="are"] [word="a" | word="an"]*, (ii) ***TERM*** *[tag="N.*"] [lemma="refer"] [word="to"]*, (iii) ***TERM*** *[tag="N.*"] [word="is"] [word="defined"] [word="as"]*, e (iv) ***TERM*** *[tag="N.*"] [word="consists"] [word="of"]*.

Quanto à possibilidade de construir SBs a partir dos padrões definitórios mais produtivos de uma área de especialidade, observamos que os resultados obtidos não derivam do agrupamento dos padrões definitórios mais produtivos do *Corpus COVID-19*, como era esperado. Tais padrões, no entanto, serviram de base para a construção das SBs. Mudanças nos atributos das SBs permitiram a construção de sintaxes que fossem capazes de extrair um maior percentual de LCs com TDs e CDs.

Por fim, com relação à possibilidade de as SBs construídas com a ferramenta *CQL* serem capazes de extrair TDs e CDs de *corpora* especializados, os resultados obtidos mostraram que as SBs complexas (por exemplo, ***TERM*** *[word="is"] [word="are"] [tag="V.*"] [word="as"])*, construídas com a *CQL* do SE a partir dos padrões definitórios mais produtivos do *Corpus COVID-19,* extraíram, em apenas uma busca, um percentual de LCs com TDs e CDs maior do que as buscas com combinação de palavras como *RNA is defined as*, ou com SBs mais simples (por exemplo, ***TERM*** *[word="defined"* | *word="presented"* | *word="known"* | *word="treated"* | *word="described"* | *word="understood"* | *word="used"]*).

Espera-se que os resultados obtidos neste estudo possam auxiliar não apenas especialistas da área médica, mas também estudantes e outros profissionais, como tradutores e lexicógrafos, a processarem de forma rápida e precisa o conhecimento expresso por meio dos termos contidos em textos especializados.

Em estudos futuros, outros padrões definitórios podem ser usados para informar a construção de SBs com a *CQL*. Da mesma forma, pode-se construir SBs a partir da identificação dos padrões definitórios mais produtivos e recorrentes em diferentes áreas de especialidade, pois os padrões definitórios variam conforme a área de especialidade e o gênero dos textos. Acreditamos que outros estudos poderão fazer uso da metodologia descrita na construção de recursos para a extração automática de CDs de *corpora* especializados.

Em uma perspectiva mais qualitativa, à semelhança do estudo realizado por Klavans e Muresan (2001), uma comparação das definições disponibilizadas

em dicionários especializados com os TDs e CDs identificados por meio das SBs trará contribuições para os estudos sobre definições.

Referências

BARROS, L. A. *Curso básico de Terminologia*. São Paulo: EdUSP, 2004.

BIBER, D., CONNOR, U., & UPTON, T. *Discourse on the Move. Using corpus analysis to describe discourse structure*. Amsterdam: John Benjamins, 2007.

BOCORNY, A. E. P., REBECHI, R. R., REPPEN, R., DELFINO, M. C. N., & LAMEIRA, V. M. A produção de artigos da área das ciências da saúde com o auxílio de *key lexical bundles*: um estudo direcionado por corpus. *D.E.L.T.A*, n., v. 1, 2021: 1-37.

CUI, H., KAN, M. Y., & CHUA, T. S. Unsupervised learning of soft patterns for definitional question answering. *Proceedings of the Thirteenth World Wide Web conference (WWW 2004)*, 2004: 90-99.

CUI, H., KAN, M. Y., & CHUA, T. S. Generic soft pattern models for definitional question answering. *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, 2005: 384-391.

DE BESSÉ, B. Le contexte terminographique. *Meta: journal des traducteurs/Meta: Translators' Journal*, n. 36, v. 1, 1991: 111-120.

DE LUCCA, J. L. Identificação de padrões recorrentes no discurso técnico e científico para a extração automática de candidatos a contextos definitórios em língua portuguesa. *Intercâmbio. Revista do Programa de Estudos Pós-Graduados em Linguística Aplicada e Estudos da Linguagem*, n. 15, 2006.

FAHMI, I., & BOUMA, G. Learning to identify definitions using syntactic features. *Proceedings of the Workshop on Learning Structured Information in Natural Language Applications*, 2006: 64-71.

FINATTO, M. J. B. Elementos Lexicográficos e Enciclopédicos na Definição Terminológica: Questões de Partida. *Organon*, n. 12, v. 26, 1998: 1-8. Disponível em < http://seer.ufrgs.br/index.php/organon/article/view/29563>. Acesso em 01 ago. 2020.

JIN, Y., KAN, M. Y., NG, J. P., & HE, X. Mining scientific terms and their definitions: A study of the ACL anthology. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 2013: 780-790.

KILGARRIFF, A.; RYCHLY, P.; SMRZ, P.; TUGWELL, D. The Sketch Engine. *Proceedings of Euralex*, 2004: 105-116.

KLAVANS, J. L., & MURESAN, S. Evaluation of the DEFINDER system for fully automatic glossary construction. *Proceedings of the AMIA Symposium*. American Medical Informatics Association, 2001: 324-328.

KLAVANS, J. L., & MURESAN, S. DEFINDER: Rule-based methods for the extraction of medical terminology and their associated definitions from on-line text. *Proceedings of the AMIA Symposium*. American Medical Informatics Association, 2000: 1049.

KOSEM, I., KOPPEL, K., KUHN, T. Z., MICHELFEIT, J., & TIBERIUS, C. Identification and automatic extraction of good dictionary examples: the case (s) of GDEX. *International Journal of Lexicography*, n. 32, v. 2, 2019: 119–137.

KOVÁŘ, V., MOČIARIKOVÁ, M., & RYCHLÝ, P. Finding definitions in large corpora with Sketch Engine. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 2016: 391-394.

MASUM, M., SHAHRIAR, H., HADDAD, H. M., AHAMED, S., SNEHA, S., RAHMAN, M., & CUZZOCREA, A. Actionable Knowledge Extraction Framework for COVID-19. Proceedings of the *2020 IEEE International Conference on Big Data (Big Data)*, 2020: 4036-4041.

PAVEL, S.; NOLET, D. *Manual de Terminologia*. Trad. Enilde Faulstich. Direção de Terminologia e Normalização. Departamento de Tradução do Governo Canadense, 2002. https://linguisticadocumentaria.files.wordpress.com/2011/03/pavel-terminologia.pdf

PEARSON, J. The expression of definitions in specialised texts: a corpus-based analysis. *Proceedings of the Seventh Euralex International Congress*, 1996: 817-824.

PEARSON, J. *Terms in context*. Amsterdam: John Benjamins, 1998.

SIERRA, G. Extracción de contextos definitorios en textos de especialidad a partir del reconocimiento de patrones lingüísticos. *Linguamática*, n. 1, v. 2, 2009: 13-37.

TRIKI, N. Elaboration paradigms in PhD theses introductions. *Deviation(s)*, 2014: 202-225.

TRIKI, N. Revisiting the metadiscursive aspect of definitions in academic writing. *Journal of English for Academic Purposes*, n. 37, 2019: 104-116.

VEYSEH, A. P. B., DERNONCOURT, F., DOU, D., & NGUYEN, T. H. A Joint Model for Definition Extraction with Syntactic Connection and Semantic Consistency. *Proceedings of the AAAI*, 2020: 9098-9105.