Programa de Pós-Graduação em

LINGUÍSTICA

IDENTIFICAÇÃO E EXTRAÇÃO SEMIAUTOMÁTICA DE CONTEXTOS DEFINITÓRIOS EM CORPUS COM VISTAS À REDAÇÃO DA DEFINIÇÃO TERMINOLÓGICA: PROPOSTA DE SISTEMATIZAÇÃO LINGUÍSTICA PARA A LÍNGUA PORTUGUESA

São Carlos
2014

DAYSE SIMON LANDIM KAMIKAWACHI

UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE EDUCAÇÃO E CIÊNCIA HUMANAS
PROGRAMA DE PÓS-GRADUAÇÃO EM LINGUÍSTICA

IDENTIFICAÇÃO E EXTRAÇÃO SEMIAUTOMÁTICA DE CONTEXTOS DEFINITÓRIOS EM *CORPUS* COM VISTAS À REDAÇÃO DA DEFINIÇÃO TERMINOLÓGICA: PROPOSTA DE SISTEMATIZAÇÃO LINGUÍSTICA PARA A LÍNGUA PORTUGUESA

Dayse Simon Landim Kamikawachi

Tese apresentada ao Programa de Pós-Graduação em Linguística da Universidade Federal de São Carlos, como parte dos requisitos para a obtenção do Título de Doutora em Linguística

Orientadora: Profa. Dra. Gladis Maria de Barcellos Almeida

São Carlos - São Paulo - Brasil

2014

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Educação e Ciências Humanas
Programa de Pós-Graduação em Linguística

Folha de Aprovação

Defesa de Tese de Doutorado da candidata Dayse Simon Landim Kamikawachi, realizada em 18/08/2014.

Comissão Julgadora:

Profa. Dra. Gladis Maria de Barcellos Almeida (UFSCar)

Prof. Dr. Leandro Henrique Mendonça de Oliveira (EMBRAPA)

Profa. Dra. Claudia Zavaglia (UNESP)

Profa. Dra. Ieda Maria Alves (USP)

Profa. Dra. Ariani Di Felippo (UFSCar)

O Relatório de Defesa assinado pelos membros da Comissão Julgadora encontra-se arquivado junto ao Programa de Pós-Graduação em Linguística.

Ao meu querido vô Mané, por me ensinar a amar as palavras. À minha amada vó Nicinha, por me mostrar que, com dedicação e afinco, é possível transformar retalhos em maravilhas.

"As palavras do ano passado pertencem à linguagem do ano passado
E as palavras do próximo ano aguardam outra voz."

— T. S. Eliot, *Little Gidding* (1942)

AGRADECIMENTOS

Esta é a página mais importante deste trabalho de Doutorado. Cheguei até aqui graças a muitas pessoas que marcaram profundamente minha trajetória acadêmica e pessoal. Destino, portanto, este pequeno espaço para registrar que esse percurso foi muito além de aulas, leituras, pesquisas e eventos.

Agradeço à CAPES, pelo apoio financeiro ao longo do Doutorado, essencial para a realização desta pesquisa.

Aos meus colegas linguistas — Maria Cristina, Claudinha, Paula Cardoso, Mari Nastri, Carol Cardoso, Maiara, Roger e Henrique Regitano —, meu carinho por compartilharmos descobertas linguísticas, trocadilhos e cookies. Vocês deixaram esta tese muito mais saborosa.

Aos professores do Departamento de Letras da UFSCar, especialmente aos Professores Nelson e Eliane, por ensinarem muito mais pela postura do que pelas palavras. À Profa. Dra. Lúcia Rino, do Departamento de Computação da UFSCar, agradeço pelos desafios propostos, pela firmeza e pelas aulas que fizeram minha cabeça explodir de ideias.

À Profa. Dra. Sandra Maria Aluísio e ao Dr. Leandro Oliveira, pelas valiosas contribuições no exame de qualificação — este último, também pelas contribuições na defesa. E às Professoras Ariani Di Felippo, Ieda Maria Alves e Claudia Zavaglia, pelas contribuições no exame final.

Às minhas amigas Isabel Bastos, Verônica Gomes e Mirian Ou, que torceram por mim, foram ouvidos quando precisei falar e braços para os meus filhos quando precisei me ausentar.

Aos meus pais, Cleide e Edilson, por serem a base firme sobre a qual construí meus sonhos — sempre com amor, apoio incondicional e confiança. À minha sogra Aparecida pelo suporte e carinho.

Às minhas pessoas favoritas: meu marido Everton e meus filhos, Anna Clara e Miguel. Obrigada por dividirem comigo os desafios e alegrias desta jornada.

Em especial, agradeço à minha orientadora, Gladis, que esteve comigo desde a Iniciação Científica até aqui. Quanta coisa aprendi com você! Ao longo desses 12 anos, tive o privilégio de aprender sobre Linguística e Terminologia, sim — mas muito mais do que isso, aprendi o que significam ética, trabalho, zelo, orientação e humildade. Devo muito do que sou a você. Ainda que os feitos científicos aqui venham a ser superados, o que sua orientação imprimiu em meu caráter permanecerá.

Por fim, agradeço a Deus, por me sustentar, me abençoar e me ensinar a cada dia. A Ele, toda honra e glória!

RESUMO

Ferramentas computacionais em Processamento de Língua Natural (PLN) são essenciais na manipulação de textos eletrônicos. Algumas ferramentas utilizadas que se podem citar são: *contadores de frequência, listas de palavras, palavras-chave e concordanciadores*. Destaca-se que esta última é a que terminólogos recorrem para visualizar e extrair contextos definitórios sobre determinado termo, os quais serão úteis na etapa da redação da definição terminológica. Ocorre que a lista de concordâncias, dependendo do termo e do tamanho do *corpus*, pode chegar muitas vezes a várias centenas de linhas, tornando a tarefa de definir extremamente morosa. Ainda que o concordanciador facilite essa tarefa humana, estudos no âmbito da Terminologia (ALARCÓN, 2009) e do PLN (KLAVANS; MURESAN, 2001) têm demonstrado que é possível desenvolver formalismo linguístico de maneira a auxiliar na geração ou enriquecimento de um sistema capaz de detectar automaticamente tais contextos. Pesquisas nessa direção têm sido realizadas para o inglês, espanhol, alemão, francês, entre outras línguas, mas para o português ainda há a necessidade de uma descrição linguística mais apurada sobre como se constituem os contextos definitórios, de modo que essa descrição possa servir de base para a construção de sistemas semelhantes para o português. Assim, esta pesquisa tem como objetivos gerais: 1) investigar padrões de contextos definitórios presentes em *corpora* de especialidades em língua portuguesa do Brasil; 2) proporcionar conhecimento linguístico que possa ser formalizado computacionalmente, a fim de integrar um sistema de extração semiautomática de candidatos a contextos definitórios; e, finalmente, 3) avaliar os resultados gerados. Na análise, foram eleitos os verbos "nomear", "conceber", "chamar", "entender", "conhecer" e "denominar" e, como *corpus* de estudo, esta pesquisa valeu-se de artigos científicos do *Banco do Português* (LAEL-PUC/SP). Foi possível realizar: 1) uma descrição quantitativa e qualitativa de cada padrão verbal definitório; 2) uma gramática local para os seis verbos, a fim de auxiliar na recuperação semiautomática de contextos definitórios; 3) uma gramática de exclusão para servir como uma *stoplist* das gramáticas locais; e 4) um conjunto de heurísticas para um classificador semiautomático de contextos definitórios. A avaliação geral apresentou precisão de 64% e cobertura de 92% na média global, o que demonstra um resultado otimista, se comparado com os demais trabalhos na literatura. Como resultado, foi possível: 1) validar a metodologia empregada a fim de estendê-la a outros padrões léxico-sintáticos; 2) obter conhecimento linguístico de modo a integrar um sistema computacional de extração semiautomática de candidatos a contextos definitórios para o português.

Palavras-chave: Contexto definitório; Definição terminológica; Processamento de Linguagem Natural; Terminologia

ABSTRACT

Computational tools in Natural Language Processing (NLP) are essential in handling electronic texts. Some of the resources commonly used are: frequency counters, word lists, keywords, and concordancers. It is noteworthy that this last tool, which terminologists use to view and extract defining contexts for certain terms, is useful in the stage of writing terminological definitions. Depending on the term and *corpus* size, the list of concordances may exceed a few hundred lines, making the task of defining the term extremely time-consuming. Yet, while concordancers facilitate the task of writing definitions, studies in terminology (ALARCÓN, 2009) and NLP (KLAVANS; MURESAN, 2001) have shown that it is possible to develop linguistic formalism that can be a substrate for the generation or enrichment of a system capable of detecting such contexts. While research in this direction has already been undertaken for English, Spanish, German, and French, among other languagesã research on Portuguese needs a more accurate linguistic explanation of defining context, in order to serve as a base for the development of similar systems for the Portuguese language. Therefore, the goals of this research are as follows: 1) to investigate the patterns of defining contexts found in technical *corpora* in Portuguese, 2) to provide linguistic knowledge which can be formalised computationally to create a system of semi-automatic extraction of candidate defining contexts, and 3) to evaluate the results generated. As the study's *corpus*, we used scientific articles from the Bank of Portuguese (LAEL-PUC), and for the analysis, the following verbs were chosen: *nomear* 'to name', *conceber* 'to conceive', *chamar* 'to call', *entender* 'to understand', *conhecer* 'to know' and *denominar* 'to denominate'. It was possible to do: 1) a quantitative and qualitative description of each verbal definitory pattern, 2) a local grammar for the chosen verbs with the purpose of aiding in the semiautomatic retrieval of definitory contexts, 3) an exclusion grammar to serve as a stoplist for local grammars and 4) a set of heuristics for a semiautomatic definitory context classifier. The evaluation of the lexical-syntactic rules of these six verbs showed 64% accuracy and 92% coverage in the global average, which represents an optimistic result in comparison to the results of previous studies. As a result, it was possible to 1) validate the methodology used, making it possible to extend it to other lexical-syntactic patterns and 2) obtain linguistic knowledge in order to integrate a semiautomatic computational system for definitory context candidates extraction for the Portuguese language.

Keywords: Definitory context; Terminological definition; Natural Language Processing; Terminology.

LISTA DE ESQUEMAS

LISTA DE FIGURAS

LISTA DE GRÁFICOS

LISTA DE QUADROS

LISTA DE TABELAS

SUMÁRIO

INTRODUÇÃO

A redação da definição terminológica (DT) se constitui numa das tarefas mais custosas e importantes no desenvolvimento de produtos terminográficos. Espera-se que, com a definição, o consulente possa ter acesso ao significado de determinado termo. A sua importância e sistematização já foram alvo de algumas pesquisas como Desmet (1990), Finatto (2002), Seppala (2007), Almeida et al (2007), Alcina e Valero Doménech (2008), Kamikawachi (2009), Wendt (2010).

Para auxiliar na redação da DT, geralmente são utilizados *corpora* textuais como fontes de compreensão sobre o termo investigado. O *corpus* é um recurso muito útil na recuperação de informações sobre dado termo em contextos reais de uso. Segundo Dubuc (1999) e a ISO (1998), os contextos onde aparecem os termos podem ser classificados em: associativos, explicativos e definitórios. Essa classificação se baseia na quantidade de informações semânticas – relacionadas aos termos – que ocorrem nesses contextos.

Especialmente, os contextos explicativos e definitórios são imprescindíveis na redação da DT, pois, a partir deles, é possível observar as propriedades semânticas do termo, tais como aspecto, constituição, origem, emprego, causa, finalidade, entre outras, as quais poderão figurar no texto da DT, dependendo da frequência com que essas informações se manifestam no *corpus*, além da consideração do público-alvo e do objetivo da obra terminográfica para a qual se está redigindo a DT.

Destaca-se que, para a visualização e recuperação dos contextos úteis para a redação da DT, normalmente se utilizam concordanciadores, ferramenta empregada no processamento de *corpus*, para listar as ocorrências de uma determinada expressão linguística que pode conter uma ou mais palavras, a qual fica em destaque, com uma quantidade definida de palavras à sua direita e à sua esquerda.

Embora o concordanciador facilite essa tarefa, estudos no âmbito da Terminologia (ALARCÓN, 2009; PINTO, 2001) e do Processamento de Língua Natural (PLN) (MURESAN e KLAVAN, 2001; XUN, 2007) têm demonstrado que, a partir de uma descrição acerca do reconhecimento dos contextos que são úteis como fontes de compreensão e de explicitação do significado do termo, é possível produzir formalismo linguístico que pode ser aproveitado para a geração ou enriquecimento de um sistema capaz de detectar tais contextos.

Pesquisas nessa direção têm sido realizadas com diversos fins e considerando padrões definitórios distintos para o inglês (MURESAN e KLAVAN, 2001), espanhol (ALARCÓN, 2009) e francês (AUGER, 1997). Para o português ainda há a necessidade de uma descrição linguística mais detalhada acerca da constituição dos contextos ricos em informação semântica, de maneira que eles possam ser recuperados por um sistema automático, auxiliando, assim, a tarefa da redação da DT.

Em parte considerável das investigações sobre o tema, geralmente privilegiaram-se regras linguísticas de extração de contextos definitórios (CD) mais genéricas do que concisas, o que possivelmente provocou uma alta cobertura e, muitas vezes, uma baixa precisão na recuperação dos CDs.

Nesse sentido, propomos, neste trabalho de doutorado, investigar padrões de CDs presentes em *corpora* de especialidades em língua portuguesa, a partir dos verbos "nomear", "conceber", "chamar", "entender", "conhecer" e "denominar", pois se constituem como importantes elementos linguísticos que ligam termos às suas respectivas definições. Daí a relevância de se fazer a descrição de seu emprego. O objetivo final da pesquisa é proporcionar conhecimento linguístico suficiente para ser integrado a um sistema de extração semiautomática de candidatos a contextos definitórios. A nossa questão de pesquisa subjacente é avaliar em que medida regras mais precisas sobre padrões verbais definitórios geram melhores resultados na extração de CDs.

O trabalho organiza-se da seguinte forma. Na primeira parte do texto, são apresentados os fundamentos da Terminologia, área na qual essa pesquisa se insere. São eles: Terminologia (seção 1), Termo (seção 2), Definição terminológica (seção 3) e Contexto terminológico (seção 4). Esses tópicos foram eleitos, pois possuem uma relação estreita com o objeto de estudo deste trabalho: o contexto definitório.

Na segunda parte, o objeto de estudo é abordado. A seção 5 é dedicada aos Aspectos práticos do contexto definitório, e a seção 6 é destinada à Revisão bibliográfica realizada sobre a identificação e extração de contextos definitórios.

Em seguida, reservamos a terceira parte do trabalho para a Metodologia utilizada na identificação de contextos definitórios (seção 7) e para a Descrição dos verbos definitórios (seção 8). Essa parte se encerra com a apresentação dos Recursos construídos para a extração de CDs, na seção 9.

Na quarta e última parte do texto, apresentamos uma breve Discussão e um apanhado dos Resultados da pesquisa (seção 10). Por fim, apresentamos as nossas Considerações finais (seção 11).

PARTE 1

FUNDAMENTOS DA TERMINOLOGIA

1. TERMINOLOGIA

> *A constituição de uma terminologia própria marca, em toda ciência, o advento ou o desenvolvimento de uma conceitualização nova, assinalando, assim, um momento decisivo de sua história. Poder-se-ia mesmo dizer que a história particular de uma ciência se resume na de seus termos específicos. Uma ciência só começa a existir ou consegue se impor na medida em que faz existir e em que impõe seus conceitos, através de sua denominação. Ela não tem outro meio de estabelecer sua legitimidade senão por especificar seu objeto denominando-o, podendo este constituir uma ordem de fenômenos, um domínio novo ou um modo novo de relação entre certos dados.*
>
> **Benveniste**

A presente seção tem como objetivo apresentar sucintamente a Terminologia, disciplina da Linguística Aplicada na qual essa pesquisa se insere. Primeiramente, abordamos um breve histórico da área e, em seguida, expomos a perspectiva da Terminologia a qual estamos filiados – a Teoria Comunicativa da Terminologia.

1.1 BREVE HISTÓRICO

A utilização de "certas palavras" em um contexto especializado é tão antiga quanto a própria existência humana, tanto é que a sistematização terminológica de uma área do conhecimento mais antiga que se tem notícia refere-se aos vocabulários temáticos monolíngues produzidos pela civilização suméria (localizada ao sul da Mesopotâmia) em aproximadamente 2600 a.C., gravados em tábuas de argila. Não coincidentemente a escrita suméria é a mais antiga língua humana conhecida[1]. De acordo com Van Hoof (1998, p. 241, *apud* BARROS, 2004, p.29), nesse vocabulário encontravam-se termos relacionados a "profissões, gado, objetos comuns e divindades".

Já em tempos menos remotos, podemos citar a descrição terminológica de Lavoisier,[2] no domínio da Química, e Linné, com seu sistema de classificação de plantas[3]

[1] SUMÉRIA. In: **Wikipédia: a enciclopédia livre**. Disponível em: <http://pt.wikipedia.org/wiki/Sum%C3%A9ria> Acesso em: jun. 2014.

[2] LAVOISIER, A.L. **Traité élémentaire de chimie**. Paris, 1789. Disponível em: <www.lavoisier.cnrs.fr/ice/ice_book_detail.php?lang=fr&type=text&bdd=koyre_lavoisier&table=Lavoisier&bookId=89&typeofbookId=6&num=0>. Acesso em fev. 2014.

[3] Disponível em: <www.botanicus.org/browse/titles>. Acesso em fev. 2014.

no século XVIII. O objetivo desses especialistas concentrava-se, sobretudo, na fixação das denominações para os conceitos científicos. O interesse dos especialistas pelos conceitos e denominações continuou no século seguinte. Devido ao desenvolvimento progressivo das ciências, os próprios especialistas manifestaram particular interesse em sistematizar regras de formação de termos para cada domínio de especialidade (CABRÉ, 1993, p. 21).

Tal interesse na sistematização terminológica pelos especialistas, motivada pelo progresso das ciências, das técnicas e da tecnologia no início do século XX, impulsionou os primeiros trabalhos que se propuseram a estabelecer a Terminologia como uma disciplina que se ocupa dos termos especializados e dos elementos que os cercam.

Nesse período, duas importantes escolas de Terminologia foram fundadas na Europa, a saber: Escola Soviética de Terminologia, por D.S. Lotte (1889-1950), e a Escola de Viena, a qual teve como cabeça o engenheiro austríaco E. Wüster (1898-1977). Embora ambas as escolas sejam consideradas precursoras da Terminologia enquanto disciplina (CABRÉ, 1993), é atribuída a Wüster a primeira exposição sistematizada de uma teoria da Terminologia, intitulada *A normalização internacional da terminologia*, que constitui sua tese de doutoramento, publicada em 1931.

Essa tese é a primeira obra de Wüster que apresenta sinais da elaboração da chamada Teoria Geral de Terminologia (TGT). Mais tarde, na década de 70, quando ele lecionou na Universidade de Viena, no curso de "Introdução à Teoria Geral de Terminologia e à Lexicografia Terminológica", Wüster apresentou os pressupostos da TGT, os quais foram publicados, após seu falecimento, no livro *Introdução* à *Teoria Geral de Terminologia e à Lexicografia Terminológica* (WÜSTER, 1998).

Em síntese, a proposta de Wüster tinha como objetivo eliminar ruídos da linguagem técnica para que esta se tornasse um instrumento eficaz de comunicação, livre de equívocos. Como forma de atingir esse propósito, a TGT apresentou um caráter estritamente normativo, o qual ignorava fenômenos inerentes à linguagem humana como sinonímia, polissemia, ambiguidade, entre outros. Ainda que esse tipo de abordagem seja válido para tratar apenas a comunicação estandardizada, ela teve seu mérito de ser a "primeira tentativa de tratar os termos de forma sistemática e coerente" (CABRÉ, 1999, p.129).

As últimas décadas do século XX foram decisivas para que ocorresse uma reflexão acerca das insuficiências da TGT, já que houve um aumento significativo das

descrições das terminologias, bem como do desenvolvimento de tecnologia adequada ao seu tratamento linguístico-computacional.

1.2 TEORIA COMUNICATIVA DA TERMINOLOGIA: PERSPECTIVA ADOTADA

A partir dos anos 80, junto às críticas à TGT, começaram a surgir perspectivas teóricas novas, as quais apresentaram como intersecção o mesmo ponto de origem, qual seja: o aporte dos conhecimentos linguísticos e de linguagem nos estudos terminológicos. Essa nova maneira de compreender as terminologias acabou por suscitar novos enfoques e/ou teorias: a Socioterminologia (GAMBIER, 1987; BOULANGER, 1991, 1995; GAUDIN 1993, 2003), a Terminologia de Base Textual (HOFFMANN, 1998; CIAPUSCIO, 2003), a Teoria Sociocognitiva da Terminologia (TEMMERMANN, 2000) e a Teoria Comunicativa da Terminologia (TCT) (CABRÉ, 1999, 2003).

Todas elas se destacam por conceber as terminologias[4] como campos de investigação pertinentes à linguagem, contudo, cada qual possui um foco específico, servindo-se de um aparato (não somente linguístico) que forneça contribuições apuradas de acordo com o que se está pesquisando.

Dentre esses paradigmas, o Grupo de Pesquisas e Estudos em Terminologia[5] (GETerm), grupo no qual se situa essa pesquisa, tem partilhado dos mesmos pressupostos teórico-metodológicos cunhados pela TCT, apresentada por Cabré (1999, 2003) e sua equipe do Instituto Universitário de Linguística Aplicada (IULA) da Universidade Pompeu Fabra (UPF) em Barcelona, Espanha. Ao contrário da TGT, a TCT articula-se baseada na valorização dos aspectos comunicativos das linguagens especializadas em detrimento dos propósitos normalizadores.

Nesse contexto, concordamos que a Terminologia seja uma disciplina linguística que deve dar conta "da descrição dos atos comunicativos especializados reais, da

---

[4] É consenso na área grafar terminologia com "t" minúsculo no sentido de "conjunto de termos", e com "T" maiúsculo quando se refere ao "campo de estudos ou disciplina".

[5] O Grupo foi criado em 1999 no Departamento de Letras da UFSCar e conta com uma equipe multidisciplinar, envolvendo linguistas, informatas e especialistas de domínio. O GETerm tem como parceiros o Núcleo Interinstitucional de Linguística Computacional (NILC), sediado no Instituto de Ciências Matemáticas e da Computação (ICMC) da USP/São Carlos; a EMBRAPA Informática Agropecuária, unidade Campinas (SP); o Instituto de Linguística Teórica e Computacional (ILTEC), sediado em Lisboa, Portugal; e o Instituto Internacional da Língua Portuguesa (IILP), sediado na cidade da Praia, Cabo Verde. O GETerm tem como objetivos: estudar conteúdos pertinentes à Terminologia/Terminografia e desenvolver pesquisas que gerem produtos terminológicos em língua portuguesa, tais como: glossários, dicionários, enciclopédias e assemelhados, que satisfaçam demandas reais.

explicação do funcionamento da terminologia enquanto manifestação da língua natural e da elaboração de aplicações diversas que satisfaçam necessidades comunicativas igualmente diversas" (CABRÉ *et al.* 1998, p. 37; 1999, p.133-134 *apud* ALMEIDA, 2003, p. 211).

Dessa forma, além de visar à elaboração de produtos terminográficos como dicionários, glossários e afins, os estudos terminológicos podem dar subsídios teórico e metodológico para áreas como Ensino de Língua com Propósitos Específicos, Tradução, Sistemas de Organização do Conhecimento (SOCs)[6], ensino de disciplinas como Medicina, Zoologia, Química, entre outros, Documentação, Jornalismo Científico, Políticas Linguísticas, Processamento de Língua Natural, etc.

As próximas seções são destinadas à discussão de três objetos inerentes aos estudos terminológicos e que se relacionam com o nosso tema de pesquisa: o termo (seção 2), a definição terminológica (seção 3) e o contexto terminológico (seção 4).

[6] Sistemas de Organização do Conhecimento (SOCs) são sistemas conceituais que representam determinado domínio por meio da sistematização dos conceitos e das relações semânticas que se estabelecem entre eles (BRASCHER; CAFÉ, 2008, p.8).

2. TERMO

Esta seção é dedicada ao principal elemento das terminologias – o termo. Primeiramente apresentamos o conceito de termo (2.1). Em seguida, abordamos alguns princípios na sua formação (2.2), e por último fazemos menção à tarefa de extração de termos em *corpus*.

2.1 DEFINIÇÃO DE TERMO

Segundo o *Termium Plus*, termo é entendido como: "Palavra (termo simples), grupo de palavras (termo composto), símbolo ou fórmula que designa um conceito particular de um campo de especialidade."[7]

A TCT tem como princípio a visão linguística sobre a linguagem especializada. Nessa abordagem, o termo ou unidade terminológica integra um determinado âmbito específico, sem perder as características próprias de qualquer unidade pertencente ao sistema linguístico das línguas naturais. Dessa forma, um termo é considerado um signo linguístico em funcionamento numa situação de comunicação especializada.

Em outras palavras, "o reconhecimento da unidade lexical como termo ou unidade terminológica se dá quando esta é definida e empregada em textos de especialidade" (KOCOUREK, 1991, p.105, *apud* BARROS, 2004, p.41). Essa afirmação tem validade no sentido de que uma unidade lexical pode ser ativada como termo se ela for portadora de um conteúdo específico dentro de um determinado domínio e, além disso, o autor aponta para a questão da definição, que pode ser um indício na identificação de termos (e suas relações) a partir de textos.

2.2 PRINCÍPIOS NA FORMAÇÃO DE TERMOS

A ISO 704 (2000, p.25-27) propõe alguns princípios para a formação de termos, a saber:

[7] Disponível em: <http://www.btb.termiumplus.gc.ca/tpv2alpha/alpha-eng.html?lang=eng&i=1&index=ptt&srchtxt=TERMO>. Acesso em jul. 2014.

a) Transparência: um termo é considerado transparente quando o conceito que ele designa pode ser inferido, ao menos parcialmente, sem uma definição. Dito de outro modo, seu significado é visível quanto ao aspecto morfológico.
b) Consistência: a terminologia de qualquer campo não deve ser uma coleção arbitrária e aleatória de termos, mas sim um sistema terminológico coerente que corresponde ao sistema de conceito.
c) Adequação: termos propostos devem ser familiares e seguir padrões estabelecidos de significado dentro de uma comunidade linguística. Formação de termos que causam confusão deve ser evitada.
d) Economia linguística: o termo deve ser o mais conciso possível. O texto da norma deixa claro que esse princípio conflita com o da precisão. Quanto maior for o número de características incluídas em um termo, maior é a precisão e a transparência do termo, contudo, seu emprego torna-se inconveniente. Assim, a norma defende que a praticidade deve reger qualquer decisão de dar preferência a um padrão de formação de termo em detrimento de outro. Por exemplo, as formas enxutas devem ser favorecidas em detrimento de formas mais longas. Observa-se ainda que um termo preciso talvez não seja desejável, por exemplo, na comunicação em determinados contextos de trabalho, enquanto que os termos complexos, mesmo formados por cinco ou seis palavras, são aceitos em publicações científicas.
e) Paradigmas derivacionais (*derivability*): a formação de termos a partir de derivação deve ser favorecida.
f) Correção linguística: um termo deve estar em conformidade com as normas morfológicas, morfossintáticas e fonológicas da língua em questão.
g) Língua nativa: apesar de os empréstimos de outras línguas serem aceitos na formação de termos, deve ser dada preferência a expressões na língua nativa.

Tais princípios podem ser vistos como diretrizes para uma comunicação eficaz, contudo, deve-se considerar que "os termos são o resultado de uma criação mais ou menos consciente" (SAGER, 1993, p.97-98), dinâmica e atualizável e, por isso, cada campo do conhecimento possui suas próprias regras de nomeação.

2.3 EXTRAÇÃO DE TERMOS EM *CORPUS*

Sager (1993) afirma que, se somos capazes de encontrar um maior número de regularidades nos modelos de denominação de elementos léxicos relacionados textualmente, então podemos chegar a:

> a) construir as regras de denominação aplicáveis a um campo temático; b) estabelecer regras para futuras designações regulares; c) e inclusive é possível que cheguemos a relacionar a motivação dos modelos de denominação à motivação mais complexa da criação dos conceitos" (tradução nossa) (SAGER, 1993, p. 98) [8].

Nesse sentido, a descrição das regras de formação de unidades terminológicas de um campo do conhecimento favorece a sua identificação e extração semiautomática em *corpus*. Podemos citar algumas iniciativas do português do Brasil, com foco nesse tipo de pesquisa, como Almeida e Vale (2010), que demonstram a eficácia da utilização de conhecimento linguístico para a extração de candidatos a termos em detrimento de métodos puramente estatísticos. Outros exemplos são a descrição da formação morfológica de termos no domínio da Nanociência e Nanotecnologia por Coleti e Almeida (2010), e Matos (2013) no domínio da Fisioterapia.

Além da investigação acerca da formação dos termos no nível da Morfologia, existem outras possibilidades de se estudar um termo ou um conjunto de termos do ponto de vista lexical ou semântico, tais como: variação linguística (SANTIAGO, 2010), sinonímia (CONTENTE e MAGALHÃES, 2011), tradução ou equivalência (BARROS, BABINI e AUBERT, 2010), rede de relações semânticas – hiperonímia, hiponímia, meronímia (MINEIRO et al., 2004), definição (DESMET, 1990), entre outros.

Na próxima seção, é apresentado um desses objetos que desempenha o papel de explicar o significado do termo inscrito como entrada no dicionário especializado – a definição terminológica.

---

[8] *a) construir las reglas de denominación aplicables a un campo temático, b) establecer reglas para las futuras designaciones regulares, c) e incluso es posible que lleguemos a relacionar la motivación de los modelos de denominación a la motivación más compleja de la creación de conceptos.* (SAGER, 1993, p.98)

3. DEFINIÇÃO TERMINOLÓGICA

Genericamente, definir é o ato de "explicar, mostrar o significado de uma palavra" (iDicionário Aulete[9]). Essa atividade humana é inerente ao processo de comunicação e, como tal, está presente em diversos níveis da nossa vida:

a) em um nível menos formal, ao definirmos um dado objeto, sentimento ou ação para um interlocutor qualquer que nos indaga ou quando notamos que nosso interlocutor ou audiência necessita de explicações sobre uma unidade lexical;
b) como também é comum em materiais didáticos, manuais, textos de divulgação científica, etc., com o propósito de esclarecer ao leitor conceitos importantes, de modo que ele avance na reflexão e na compreensão do domínio descrito;
c) até chegar ao seu alto nível de formalização e comprometimento, quando inserida em verbetes de dicionários, glossários e afins.

Uma definição natural (níveis "a" e "b") e uma definição dicionarizada (representada pelo item "c") se distinguem, sobretudo, pelo seu modo de apresentação. Enquanto nas duas primeiras é permitido o uso de uma metalinguagem como "é definido como", "pode ser considerado", "é designado como", "significa", ou seja, expressões que evidenciam que se trata de um contexto definitório, a definição dicionarizada se caracteriza por determinadas convenções e princípios que regem sua estrutura e apresentação, balizados de acordo com a perspectiva teórica e metodológica adotada na construção do artefato linguístico.

Além disso, como poderemos ver adiante, a definição dicionarizada se serve (ou deveria se servir) de fontes fiáveis de textos autênticos que apresentam a unidade lexical em uso, para auxiliar na construção de seu significado no verbete. Em contrapartida, ao expressarmos por escrito ou oralmente a definição de uma unidade lexical de forma mais ou menos espontânea, muitas vezes nos apoiamos em informações semânticas já registradas em definições dicionarizadas.

Larivière (1996, p.409) discrimina os três tipos de definições dicionarizadas em:

[9] Disponível em: <aulete.uol.com.br/>. Acesso em fev. 2014.

> **Definição lexicográfica (DL):** utilizada nos dicionários de língua e nos dicionários enciclopédicos, propõe-se a explicitar os significados distinguindo os sentidos e os empregos dos signos (ou palavras) de uma língua;
> **Definição enciclopédica (DE):** utilizada nas enciclopédias e nos dicionários enciclopédicos, propõe-se a fornecer um conjunto de conhecimento sobre uma coisa;
> **Definição terminológica (DT):** utilizada nos vocabulários especializados, propõe-se a caracterizar (delimitar e distinguir de outras noções) as noções denominadas por um termo e que representam uma coisa no interior de um sistema organizado. (Tradução nossa). (LARIVIÈRE, 1996, p. 409) [10]

Para Bessé (1990, p. 254, *apud* LARIVIÈRE, 1996, p.413), a diferença entre a definição enciclopédica para a terminológica reside no fato de que esta última é interrompida após ter fornecido toda a informação que permite situar e diferenciar o conceito dentro do sistema conceitual. Já a definição enciclopédica "fornece toda a informação possível sobre o conceito, independentemente da sua relevância" (COUTO, 2004, p.16).

Portanto, é de se esperar que um mesmo termo seja apresentado de formas distintas, a depender do caráter da obra – terminológica, enciclopédica ou lexicográfica. E essas diferenças certamente influenciam na constituição do texto definitório.

Tal como ocorre na produção da escrita nos diversos gêneros textuais (materiais didáticos, manuais, textos de divulgação científica, etc.), a etapa que corresponde à definição exige certo planejamento anterior à sua redação, principalmente levando-se em consideração o público-alvo e a finalidade da obra em questão. Isto é, da mesma forma que o autor de um manual ou livro didático tem como objetivo transmitir um determinado assunto ou fazer com que o leitor execute uma dada tarefa e, para tal, o autor necessariamente apresenta uma linguagem que seja adequada a esse usuário em termos estruturais e de conteúdo, o mesmo acontece na redação da DT, em que o redator prevê as necessidades dos consulentes em potencial e, a partir disso, adequa o texto definitório.

A seguir, a definição terminológica é detalhada quanto à sua natureza, princípios de redação e às etapas seguidas na sua elaboração.

---

[10] ***la définition lexicographique (DL):*** *utilisée dans les dictionnaires de langue et les dictionnaires encyclopédiques, qui se propose d'expliciter des signifiés en distinguant les sens et les emplois des signes (ou mots) d'une langue;* ***la définition encyclopédique (DE):*** *utilisée dans les encyclopédies et les dictionnaires encyclopédiques, qui se propose de fournir un ensemble de connaissances sur une chose;* ***la définition terminologique (DT):*** *utilisée dans les vocabulaires spécialisés, qui se propose de caractériser (i.e. de délimiter et de distinguer des autres notions) des notions dénomimées par un terme et représentant une chose à l'intérieur d'un système organisé.* (LARIVIERE, 1996, p.409)

3.1 DEFINIÇÕES DA DEFINIÇÃO TERMINOLÓGICA

A definição terminológica tem sido consideravelmente debatida dentro dos organismos normalizadores do trabalho terminológico, por terminólogos consagrados e grupos de pesquisas que têm como foco a construção de produtos terminográficos. Nesta subseção, serão apresentadas algumas perspectivas sobre o tema em relação à sua natureza.

Segundo a ISO 1087-1 (2000, p. 6) a definição é "uma representação de um conceito por uma declaração descritiva que serve para diferenciá-lo de conceitos relacionados." (tradução nossa)[11]. A norma apresenta dois grandes tipos de definição: a intensional[12] e a extensional[13]. A primeira indica "o conceito superordenado, imediatamente superior ou um ou dois níveis acima, seguido das características que distinguem o conceito dos outros conceitos" (tradução nossa) (ISO 704.1, 2000, p. 15)[14]. Já as definições extensionais são constituídas por uma "lista dos conceitos subordinados, designando os objetos que compõem a extensão do conceito" (tradução nossa) (ISO 704.1, 2000, p.17)[15]. Esses são os dois grandes tipos de definição usados no trabalho terminográfico, os quais tiveram sua origem na Filosofia, juntamente com outras variedades de definição (COUTO, 2004, p.11).

No *Vocabulaire systématique de la terminologie*, organizado por Boutin-Quesnel et al. (1985, p. 27), a *definição* é entendida como um enunciado que descreve uma noção e que permite diferenciá-la das outras noções no interior de um sistema nocional.

Para *O Pavel: Curso Interativo de Terminologia*, "a definição é um breve enunciado lexicográfico que fornece as características essenciais e distintivas de um conceito e indica o lugar do conceito em um sistema conceitual".

---

[11] *definition:* ɾepresentation *of a concept by a descriptive statement which serves to differentiate it from related concepts.* (ISO 1087-1, 2000, p.6)

[12] A ISO 704.1 (2000, p. 16) traz o seguinte exemplo de definição intensional: *pencil case: container designed to hold and carry pencils and other writing instruments* ("estojo: recipiente projetado para armazenar e transportar lápis e outros instrumentos de escrita", tradução nossa).

[13] A ISO 704.1 (2000, p.17) traz o seguinte exemplo de definição extensional: *threatened species: critically endangered species, endangered species or vulnerable species.* ("espécies ameaçadas: espécies criticamente ameaçadas de extinção, espécies em risco de extinção ou espécies vulneráveis", tradução nossa).

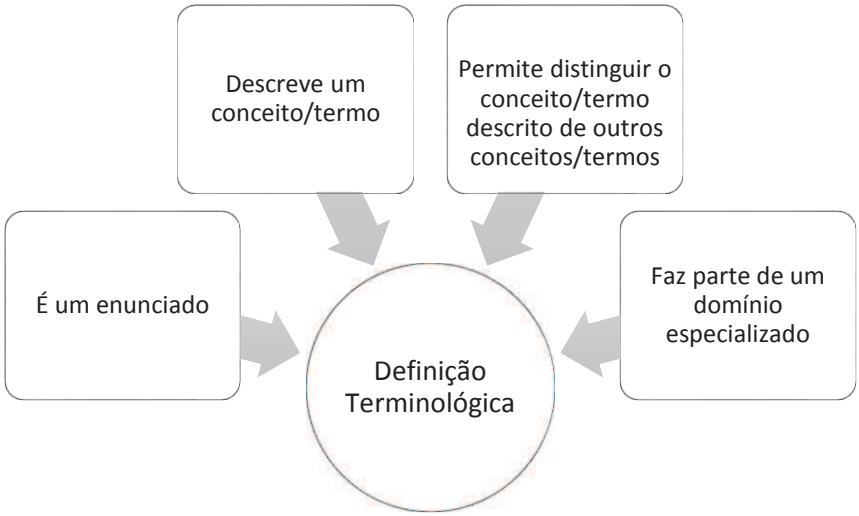
[14] *Intensional definitions* shall indicate the *superordinate concept*, either immediately above or at a higher level, followed by the *characteristic(s*) that distinguish the *concept* from other *concepts*. (ISO 704.1, 2000, p. 15)

[15] *(...) the definition can be formulated as a list of the subordinate concepts, in only one dimension, which correspond to objects making up the extension of the concept.* (ISO 704.1, 2000, p. 17)

Já o documento elaborado pelo *Office québécois de la langue française* (VÉZINA, R. et al., 2009, p.7), o qual tem como propósito servir de orientação na redação da definição terminológica, postula que a definição pode ser compreendida como uma "'carteira de identidade' de um termo dentro de um conjunto terminológico, pertencendo a um domínio particular." Ela "faz o elo entre uma denominação e um conteúdo conceitual."[16] (tradução nossa).

As quatro fontes citadas fazem referência a quatro importantes propriedades da definição terminológica, elencadas no Esquema 1.

Esquema 3 - Propriedades da definição terminológica



Fonte: Elaborado pela autora.

É possível concluir que a definição terminológica desempenha um importante papel na comunicação especializada, uma vez que ela deve dar conta de transmitir o significado de um termo e, além disso, de diferenciá-lo dos demais termos que estão inseridos no mesmo domínio especializado.

Para além dos documentos citados, importantes terminólogos também abordaram o caráter conceitual da definição terminológica. Neste trabalho, sinteticamente são expostos os pontos de vista de Wüster (1998), Rey (1979), Felber (1984), Desmet (1990), Sager (1993), Cabré (1993), Dubuc (1999) e Finatto (2001).

[16] *La définition constitue en quelque sorte la 'carte d'identité' d'un terme au sein d'un ensemble terminologique, lui-même appartenant à un domaine particulier. On peut considérer qu'une définition fait le lien entre une dénomination et un contenu conceptuel.* (VÉZINA, R. et al, 2009, p.7)

O precursor dos estudos terminológico, Wüster, estipula que uma definição "é a descrição de um conceito por meio de conceitos conhecidos, expressos geralmente por palavras" (tradução nossa)[17] (WÜSTER, 1998, p. 65).

O lexicógrafo Alain Rey (1979) afirma que, empiricamente, "a definição terminológica é um compromisso entre definição lexicográfica e descrição enciclopédica", destinada a: 1) "melhorar o uso dos nomes para lhes permitir funcionar como termos"; 2) "evocar o modo de constituição das classes de seres e o funcionamento dos esquemas conceituais" (tradução nossa) (REY, 1979, p. 43).

Seguidor de Wüster, Felber considera que a "definição é a descrição de um conceito por meio de outros conceitos conhecidos, realizada através de palavras ou termos, determinada pela posição do conceito dentro do sistema conceitual." (tradução nossa) (FELBER, 1984, p. 160). Desse modo, a elaboração da definição terminológica deve ter como base o lugar do conceito dentro de um sistema conceitual e as relações estabelecidas entre ele e os demais conceitos desse mesmo sistema.

Para Desmet (1990, p. 6), "a definição é compreendida como classificadora, hierarquizante e estruturante". A autora defende ainda que "não é possível a utilização de uma linguagem técnica ou científica sem definições" (DESMET, 1990, p.5).

No capítulo referente à definição do *Manuel pratique de terminologie* (DUBUC, 1999), o autor afirma que o objetivo da definição é "dar uma imagem mental exata de um conceito (...) que permita sua identificação, baseando-se em seus traços essenciais" (tradução nossa) (DUBUC, 1999, p. 119)[18].

Já Sager (1993) postula que a definição terminológica, enquanto produto, "é uma descrição linguística de um conceito, constituída pela enumeração de um conjunto de características que dão conta do significado do conceito" (tradução nossa) (SAGER, 1993, p. 68)[19]. O autor faz referência ao tipo clássico de definição composto pelo gênero próximo e diferença específica (GPDE), que se constituiu como referência para a elaboração de definição desde os estudos filosóficos, passando pela Terminologia Clássica, até alcançar lugar de destaque nos estudos terminológicos nos dias de hoje.

---

[17] *en su sentido amplio, una definición es la descripción de un concepto mediante conceptos conocidos, expresados generalmente por medio de palabras*. (WÜSTER, 1998, p. 65).

[18] *dar una imagen mental exacta de un concepto (...) que permita su identificación, basándose en sus rasgos esenciales."* (DUBUC, 1999, p. 119)

[19] *es una descripción linguística de un concepto. Basada en el listado de un número de características que transmiten el significado del concepto."* (SAGER, 1993, p. 68)

De acordo com o pesquisador, as definições especializadas descrevem um conceito dentro de um campo temático especializado: "Uma definição terminológica oferece uma identificação única de um conceito somente com referência ao sistema conceitual do qual faz parte e classifica o conceito dentro desse sistema" (tradução nossa) (SAGER, 1993, p. 68)[20].

Cabré (1993) também tratou da definição terminológica e a definiu como "uma expressão normalmente complexa equivalente semanticamente ao termo que define" (CABRÉ, 1993, p. 312), em referência exclusiva a um domínio específico. Nesse sentido, a pesquisadora defende que "um conceito pode ser representado quer por uma definição quer por uma ilustração" (CABRÉ, 1999, p.104), sendo que "a primeira possibilidade se dá por meio de fórmulas linguísticas, enquanto que a segunda se dá por meio de unidades icônicas a partir da reprodução da ideia que os indivíduos têm de uma classe de objetos" (tradução nossa).

Por fim, fazemos menção ao trabalho de doutorado defendido por Finatto (2001), denominado *Definição terminológica: fundamentos teórico-metodológicos para sua descrição e explicação*. A pesquisa oferece um renovado paradigma teórico e abre novos olhares para a definição terminológica, a partir da observação das condições de subjetividade reveladas do texto definitório, utilizando, para isso, os conceitos fundamentais da semântica enunciativa cunhados por Émile Benveniste.

Finatto (2001) acredita que é essencial considerar a definição terminológica como um objeto linguístico multifacetado, construído pelo indivíduo-autor e pela coletividade que ele representa. Dessa forma, na perspectiva da autora, a definição terminológica é "dotada de características que a fazem exceder aqueles limites mais usuais ou tradicionais de um objeto lógico-categorial" (FINATTO, 2001, p.14).

Diante da posição dos estudiosos acerca do conceito de definição, podemos perceber que devido ao fato de a reflexão sobre sua constituição ter sido originada no âmbito da Filosofia, as abordagens mais linguísticas como em Cabré (1993, 1999), Sager (1993) e Desmet (1990) apresentam vestígios desse pensamento quanto à elaboração da definição. Por outro lado, há também esforços em compreender a definição como um texto especializado que reflete o "entorno de significação" de dado termo e da área do

[20] *Una definición terminológica ofrece una identificación única de un concepto sólo con referencia al sistema conceptual del que forma parte y clasifica el concepto dentro de ese sistema.* (SAGER, 1993, p. 68)

conhecimento do qual faz parte, como também da percepção do objeto linguístico pelo seu sujeito-redator (FINATTO, 2001).

A subseção 3.2 é destinada a apresentar algumas questões acerca da prática da elaboração da definição terminológica.

3.2 ELABORAÇÃO DA DEFINIÇÃO TERMINOLÓGICA

Cabe à equipe gestora do projeto de desenvolvimento de produtos terminográficos tomar decisões acerca do protocolo de elaboração da definição terminológica. Para isso, existem alguns critérios importantes e inter-relacionados que guiam a tarefa. São eles: princípios e convenções na redação da definição terminológica (3.2.1), tipologia de definições (3.2.2), adequação ao público-alvo (3.2.3) e reprodução *versus* redação da definição (3.2.4).

3.2.1 Princípios e convenções na redação da definição terminológica

Juntamente com o conceito da definição terminológica, os documentos que tratam do assunto também apresentam regras ou convenções quanto à redação do texto definitório com o objetivo de que este cumpra seu papel de esclarecer o significado de um termo de modo satisfatório.

Podemos observar que as convenções são comuns em boa parte dos trabalhos que tratam do tema da definição terminológica. Muito provavelmente, esse fato se deva a postulados filosóficos que se detiveram em regras de como elaborar uma boa definição. Por exemplo, COUTO (2004, p. 21) afirma que "a obra *Tópicos* de Aristóteles serviu de base para a fundamentação das regras de redação da definição na ISO 704".

Como as convenções se assemelham consideravelmente, aqui são mencionadas as descritas em Cabré (1993, p. 312-313)[21]. A autora as divide em três grupos: I. Adequações gerais; II. Adequações específicas e III. Expressão.

[21] O texto original foi traduzido e adaptado por nós.

I. Adequações gerais

As definições devem:

a) ser verdadeiras;

b) permitir a distinção dos outros conceitos;

c) reconhecer as dimensões pertinentes de cada campo de especialidade;

d) situar-se na perspectiva do campo nocional a que pertence o conceito;

e) adequar-se às finalidades do trabalho de que faz parte.

II. Adequações específicas

As definições devem:

a) acoplar-se ao sistema de definições de um campo concreto, partindo da estrutura prévia deste campo;

b) explicitar todas as características (traços) essenciais de cada conceito, de acordo com a estrutura nocional estabelecida;

c) refletir as relações sistemáticas que cada conceito mantém com os demais conceitos do mesmo campo;

d) incluir todas as características que, ainda que não sejam essenciais, são importantes para uma descrição completa do conceito;

e) apresentar os traços dos conceitos de forma sistemática, especialmente quando se trata de conceitos que pertencem a um mesmo campo do mapa conceitual.

III. Expressão

As definições devem:

a) ser expressas corretamente;

b) ser formalmente adequadas, de acordo com as normas de construção de definições;

c) utilizar a expressão adequada aos destinatários do trabalho;

d) constar de uma só oração, evitando pontos internos;

e) respeitar os princípios da Lexicografia no que diz respeito a sua apresentação formal, que se resumem aos seguintes pontos:

- o descritor inicial deve ser da mesma categoria gramatical do termo descrito;
- devem-se utilizar palavras conhecidas pelos usuários em geral;
- ao utilizar vocábulos mais específicos, eles também têm de ser entradas;

- a definição não deve ser circular;
- evitar a negação, a utilização de paráfrases desnecessárias e de fórmulas metalinguísticas.

Fazemos referência também a algumas orientações interessantes que devem ser levadas em consideração na elaboração da definição, presentes em Castillo (1997, p. 79-90[22]):

a) a redação das definições deve ser clara e concisa;
b) não se devem expressar critérios de valor não fundamentados devidamente sobre bases estritamente técnicas;
c) o estilo do texto definitório deve ser impessoal;
d) se for necessário citar definições divergentes para a adequada compreensão do termo, o ideal é que isso seja feito com total imparcialidade, sem incorrer em análises polêmicas;
e) a ordem das acepções pode ser estabelecida a partir de um dos dois princípios: i) do sentido mais frequente ao menos frequente ou do geral para o mais específico, ii) do sentido historicamente mais antigo ao mais recente;
f) uma definição deve ser autossuficiente, completa em si mesma e não dependente de nenhum outro elemento;
g) em virtude dessa autossuficiência, deve-se considerar inadequada num texto definitório a repetição do termo-entrada;
h) o texto definitório deve ser preferencialmente redigido no singular;
i) as definições devem ser sintaticamente equivalentes ao termo definido, posto que todo verbete de dicionário é a expressão de uma relação de igualdade estabelecida entre o termo definido e sua definição, portanto, se o termo-entrada tem valor substantivo, adjetivo ou verbal, a definição deve corresponder à categoria gramatical de substantivo, adjetivo ou verbo, respectivamente;
j) deve-se evitar utilizar na definição termos desconhecidos, mas se isso não for possível, os termos utilizados no texto definitório devem-se tornar entradas de outros verbetes;

[22] O texto original foi traduzido e adaptado por nós.

k) para economizar esforços na elaboração da DT, procede-se do mesmo modo que numa pesquisa léxica: aproveitam-se as existentes em dicionários técnicos e outras fontes confiáveis. Dessas definições, elimina-se toda informação que não seja imprescindível para a identificação do conceito e, a partir delas, elabora-se uma redação adequada aos padrões terminológicos.

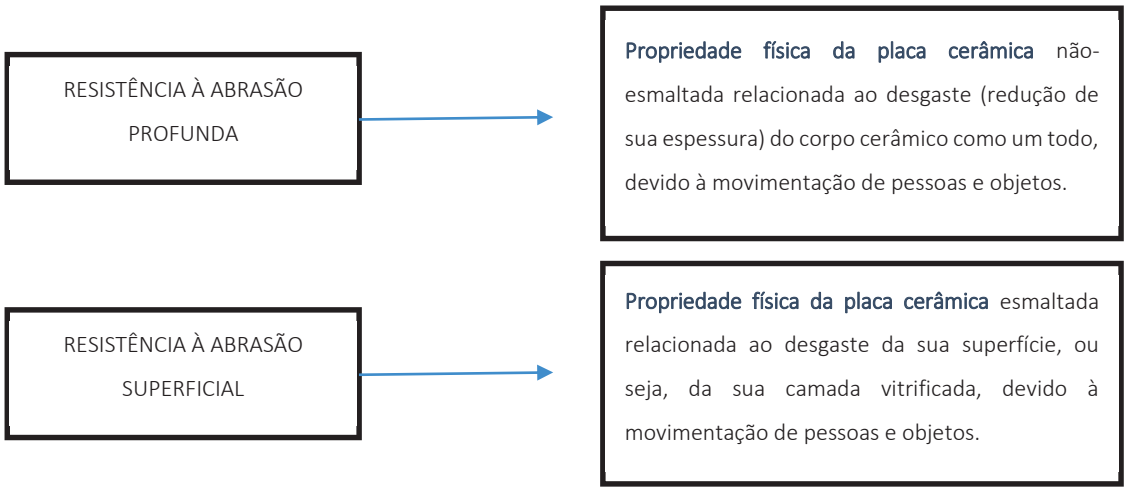
Esses princípios e regras servem de orientação às equipes que têm como alvo a elaboração de definições em projetos terminológicos. Pela nossa experiência na elaboração de diversos dicionários no âmbito do GETerm, podemos afirmar que muitos deles são úteis para a formação da equipe de trabalho, especialmente quando os membros são novatos na área dos estudos lexicográficos ou terminológicos. Contudo, na atividade da redação da definição, termos podem apresentar determinadas particularidades que, por sua vez, necessitam romper com alguma convenção, com o propósito de fornecer ao usuário o seu significado da melhor forma possível. E esse deve ser o princípio máximo na elaboração da definição terminológica – a adequação ao público, pois subordiná-lo a qualquer outra coisa seria perder o sentido do trabalho terminológico comunicativo, descritivo e de base linguística.

3.2.2 Tipologia de definições

A definição intensional ou definição por gênero próximo e diferença específica (GPDE) é considerada o tipo clássico de definição. Como já mencionamos anteriormente, esse tipo se constituiu como referência para a elaboração de definição desde os estudos filosóficos, passando pela Terminologia Clássica, até alcançar lugar de destaque nos estudos terminológicos nos dias de hoje.

A definição que segue o modelo GPDE apresenta-se da seguinte forma: a DT é encabeçada pelo gênero próximo (GP), ou seja, um hiperônimo (ou termo-pai), o qual geralmente é recuperado observando-se o termo superordenado na ontologia ou na estrutura conceitual. Imediatamente, segue(m)-se a(s) diferença(s) específica(s) (DE), características que particularizam determinado termo em relação aos demais termos que possuem o mesmo GP. Apresentamos no Esquema 2 um exemplo contendo duas DTs, retiradas do *Glossário de revestimento cerâmico* (ALMEIDA et al., 2011), que apresentam o mesmo GP.

Esquema 4 – Definições terminológicas por GPDE



Fonte – elaborado pela autora.

Além desse tipo de definição e da definição por extensão proposta pela ISO (2000, p.15),[23] Sager (1993, p. 2-73[24]) apresenta uma tipologia variada e flexível em comparação à referida norma:

a) **Definição mediante análise**. Exemplo: gengivite = inflamação da gengiva.
a) **Definição mediante sinônimos.** Exemplo: margarida-do-campo = margarida-anual (*Novo Dicionário Eletrônico Aurélio*, versão 6.0**,** 2009).
b) **Definição mediante paráfrase**. Exemplo: brancura = qualidade do que é branco (*Novo Dicionário Eletrônico Aurélio*, versão 6.0**,** 2009).
c) **Definição mediante síntese** (mediante relação identificativa, mediante descrição). Exemplo: margarida = em algumas máquinas de escrever e impressoras, disco com palhetas dispostas radialmente, na extremidade das quais há tipos em relevo para a impressão de caracteres (*Novo Dicionário Eletrônico Aurélio*, versão 6.0**,** 2009).
d) **Definição mediante implicação** (mediante o uso da palavra em um contexto explicativo). Exemplo: diagnóstico = faz-se um diagnóstico quando se identificam certos sintomas como característicos de uma condição especial.

[23] Tipologia de definição já apresentada na subseção 3.1. O item "f" coincide com a definição por extensão proposta pelo documento da ISO (2000, p.15).
[24] Traduzido e adaptado por nós.

e) **Definição mediante denotação** (mediante a enumeração de exemplos, mediante extensão). Exemplo: cão = fila, buldogue, *poodle*, *pit bull*, etc.
f) **Definição mediante demonstração** (definição ostensiva). Exemplo: desenhos, fotografias, referência situacional.

Existe a possibilidade também de integrar dois tipos em uma mesma definição, como por exemplo: mediante **análise e descrição**; mediante **sinônimo e descrição**; mediante **sinônimo e análise**, entre outros (ALMEIDA, SOUZA e PINO, 2007).

A escolha do tipo de definição que será empregada no projeto terminográfico está diretamente relacionada com a natureza do termo, ou seja, aspectos do termo como sua classificação e constituição morfológica interferem na decisão sobre qual tipo definitório adotar. Por exemplo, um termo que é classificado como adjetivo poderá ser definido mediante sinônimo e implicação, mas dificilmente será definido mediante o tipo clássico de definição GPDE.

Outro aspecto que interfere no emprego do tipo de definição é adequar a definição ao público-alvo do produto terminográfico, de modo a não deixá-la incompreensível e nem tão superficial. Essa adequação diz respeito tanto ao vocabulário que está sendo empregado no texto definitório como também à escolha do tipo definicional.

A subseção 3.2.3 aborda esse aspecto da elaboração da definição terminológica.

3.2.3 Adequação ao público-alvo

Avaliar as necessidades do potencial público-alvo do produto terminográfico é uma tarefa essencial no planejamento da adequação da redação da definição, todavia, segundo os autores Atkins e Rundell (2008), essa questão tem sido pouco investigada em comparação com aspectos mais teóricos da definição.

Assim como um mesmo tema pode ser dirigido a públicos distintos com *background* e interesses diferentes, o mesmo pode ocorrer com a consulta aos dicionários –diferentes perfis podem buscar o significado de um termo, porém com expectativas distintas. Sager (1993, p.81) distingue quatro classes de potenciais públicos-alvo: "terminólogos, tradutores, especialistas e usuários não especializados."

Cada classe citada tem suas próprias necessidades na consulta à definição. Por exemplo, o terminólogo se serve da definição existente para criar sua própria definição ou ainda para a alocação do termo definido numa ontologia, numa estrutura conceitual ou numa taxonomia. No caso dos tradutores, a definição pode ajudá-lo a compreender se o termo empregado é de fato a equivalência mais próxima do termo no original. Por sua vez, os especialistas que têm conhecimento razoável na sua área lançam mão das definições para sanarem eventuais dúvidas e talvez esse seja o público mais rigoroso quanto ao conteúdo da definição, pois conforme apontaram Almeida, Souza e Pino (2007), o público especializado é o que espera do dicionário definições mais exaustivas e menos enxutas. Já o usuário não especializado, que pode ser um estudante ou ter qualquer outra ocupação diferente das citadas anteriormente, no nosso ponto de vista, é o perfil com características mais difíceis de prever e que, por isso, pode ter expectativas diversas, pois pode desejar uma explicação mais superficial sobre o termo, caso não tenha nenhum ou pouco conhecimento da área de conhecimento descrita, ou pode, como afirma Sager (1993, p.81), "necessitar de uma explicação de caráter enciclopédico," posto que qualquer outro tipo de definição não lhe seria útil.

Uma definição altamente especializada muito provavelmente será de difícil compreensão para um novato no domínio; já uma definição muito didática e simplificada pode ser considerada incompleta e ineficaz para um público mais especializado. Portanto, é essencial que a equipe de trabalho considere a importância de se ter clareza para quem as definições estão sendo elaboradas e, dessa forma, obter definições homogêneas quanto à quantidade de informação veiculada, à disposição dessa informação no texto definitório e ao vocabulário empregado.

3.2.4 Reprodução versus redação da definição

Há projetos terminográficos em que a opção é trazer a definição extraída de textos autênticos, tal qual ocorre no uso real pelos especialistas ou para os especialistas. Nesse modo de apresentação do verbete, o termo é seguido de uma ou mais definições elaboradas geralmente por autores consagrados do domínio em questão, seguida(s) da referência da fonte. Quando necessário, é ainda fornecida uma definição criada para o termo dicionarizado, com o propósito de esclarecer algum aspecto do seu significado que não tenha ficado explícito com as outras definições.

Podemos citar algumas vantagens nessa abordagem de trabalho para a equipe desenvolvedora e para o usuário final, tais como: diminuição considerável do tempo dispensado na confecção do trabalho terminográfico, já que a etapa da redação da definição, devido à complexidade da tarefa, leva certo tempo; aumento de percepção da área de estudo de forma mais nítida ao ter acesso a definições originais redigidas pelos próprios especialistas, com suas próprias escolhas lexicais e de eleição de quais informações semânticas sobre o termo serão privilegiadas; observação do comportamento morfossintático do termo na própria definição; e por último, recuperação rápida da fonte textual de onde se extraíram a definição e o termo.

Por outro lado, há desvantagens como: definições pouco didáticas ou que apresentam falhas na sua estrutura redacional; provável utilização de fontes textuais nem sempre oriundas de entidades competentes e confiáveis e, geralmente, a não utilização de um *corpus* especializado adequado à proposta da obra terminológica.

Sobre a necessidade de elaborar definições terminológicas em bancos terminológicos, Sager (1993, p. 81-82) afirma que "não é necessário e nem útil definir termos que já foram definidos exaustivamente, como substâncias químicas, minerais, plantas, métodos de análise científica, medidas, entre outros." Nessa perspectiva, por exemplo, em um dicionário destinado aos termos da Fisioterapia, não se tornariam entradas termos do corpo humano, como braço, cabeça, quadril, já que são termos compreendidos por todos. Em contrapartida, em um dicionário de Mineralogia e Geologia, seria frustrante para o usuário abrir a obra e não ver o termo *grafite* ou *diamante* como entrada.

Isso posto, acreditamos que essa questão mencionada por Sager (1993) sobre o uso de definições extraídas de textos autênticos deve ser discutida, levando-se em conta o repertório terminológico, o objetivo e o público-alvo da obra terminográfica.

Concluímos a seção 3.2 reiterando a importância de se examinar cada um dos três pontos abordados previamente à elaboração do conjunto de definições, de forma a se obter definições homogêneas e que atendam aos objetivos do trabalho. Na seção 3.3, são elencados os recursos empregados na etapa da redação da DT, os quais podem dar subsídios também aos critérios utilizados para a sua elaboração.

3.3 RECURSOS PARA A REDAÇÃO DA DEFINIÇÃO TERMINOLÓGICA

Nessa seção, são apresentados os principais recursos utilizados na elaboração e validação da definição terminológica. Tais recursos não são elementos obrigatórios em um projeto terminológico, contudo, são extremamente úteis em várias atividades, dentre elas, na identificação e seleção de informação semântica que poderá constar da definição.

Os recursos são descritos obedecendo à sequência: ontologia (3.3.1), especialista de domínio (3.3.2), base definicional (3.3.3) e *template* (3.3.4).

3.3.1 Ontologia

Com o conjunto de termos em mãos, o passo seguinte é a alocação desses termos em uma ontologia ou estrutura conceitual, forma de organização do conhecimento que permite a visualização de diversas relações (como hiperonímia, hiponímia e co-hiponímia, etc.) entre os termos/conceitos que compõem o domínio com o qual se está trabalhando. A ontologia cumpre a função de explicitar um conjunto de termos/conceitos e a forma como eles se relacionam. Sua importância reside no fato de que ao alocarmos um termo/conceito em um determinado ponto da ontologia, assumimos que o definiremos de acordo com o campo ao qual ele foi destinado. É importante frisar que, independente do domínio, a ontologia pode se organizar de diferentes maneiras, dependendo da perspectiva da pessoa ou da equipe encarregada da sua elaboração, pois como afirmam Friedman Noy e McGuinness (2001), "uma ontologia é um modelo de realidade do mundo que não é fixo, mas dinâmico e os conceitos da ontologia devem refletir essa realidade." (tradução nossa)[25]

A subseção 3.3.2 aborda a importância do especialista de domínio na construção do texto definitório em projetos terminológicos.

3.3.2 Especialista de domínio

Em um cenário ideal de criação de produtos terminográficos, a equipe deverá ser multidisciplinar – formada por, no mínimo, um terminólogo e um especialista de domínio. Isso porque ambas as formações permitem que cada qual contribua com diferentes

[25] *An ontology is a model of reality of the world that is not fixed, but dynamic and the concepts in the ontology must reflect this reality"* (FRIEDMAN NOY e MCGUINNESS, 2001).

elementos do conhecimento. Se o terminólogo tem habilidades linguísticas e da própria Terminologia, enquanto disciplina, para organizar termos e definições de uma área qualquer, o especialista de domínio tem consigo o conhecimento conceitual da área descrita. Os especialistas de domínio devem ser "profissionais competentes, reconhecidos como autoridades no seu campo de conhecimento" (DUBUC e LAURISTON, 1997, p. 87).

Ambos os profissionais participam do processo terminográfico de modo complementar e com divisão clara de tarefas. Podemos citar o contexto de trabalho da cooperação entre o GETerm e a EMBRAPA, no qual vários domínios[26] têm sido sistematizados, levando-se em conta o conhecimento de linguistas/terminólogos e especialistas de domínio.

Os especialistas são importantes na indicação das melhores fontes textuais para a compilação do *corpus*, validação de lista de termos, cooperação na elaboração da ontologia e validação da definição terminológica. Por sua vez, os terminólogos são profissionais aptos a compilar o *corpus* terminológico, extrair listas de candidatos a termos, elaborar a ontologia e redigir a definição.

Contudo, pode ocorrer troca de papéis. No caso específico da etapa da redação da definição, o especialista pode redigi-la, e o linguista fazer a validação. Não consideramos isso como uma inadequação metodológica, uma vez que, da mesma forma que o terminólogo aprende com o *corpus* e com o especialista, este pode, da mesma forma, obter conhecimento junto ao terminólogo sobre aspectos formais e de conteúdo da sua redação.

A compilação mais ágil e menos morosa de grandes *corpora* permite, sobretudo ao terminólogo, ter acesso ao conhecimento de forma dinâmica, o que contribui com a sua formação mais especializada do domínio, deixando ultrapassada a afirmação de Bessé (1997, p.73) de que a maioria das definições terminológicas é escrita por especialistas de domínio porque terminólogos e, por fortes razões, lexicógrafos, não têm sempre conhecimento suficiente do domínio para estarem aptos a escrever definições apropriadas.

Assim, para além da divisão das tarefas entre os terminólogos e especialistas de domínio, é essencial que haja cooperação entre eles para que o produto desenvolvido

[26] Intensificação Agropecuária (INTAGRO), Recursos Hídricos e Geoinformação Espacial (GeoInfo) são áreas que estão sendo investigadas quanto aos seus termos e conceitos pela Embrapa Informática Agropecuária, unidade Campinas, SP.

resulte em um bom artefato. Por fim, finalizamos essa subseção com uma citação que expressa de forma singular a relação do especialista com o terminólogo:

> Uma espécie de simbiose deve se desenvolver entre estes dois tipos de especialistas. O trabalho do terminólogo atesta a validade linguística dos dados; o conhecimento do especialista de domínio garante que a investigação permaneça em conformidade com a realidade da área temática. (DUBUC e LAURISTON, 1997, p.87) (tradução nossa).[27]

3.3.3 Base definicional

A base definicional[28] é outro recurso importante que auxilia na redação da definição. Trata-se de um banco de contextos definitórios e explicativos[29] referentes ao conjunto de termos, compilados de diversas e variadas fontes, tais como o próprio *corpus* da pesquisa, como também outras fontes não incluídas no *corpus*, mas que podem se revelar fontes úteis para a obtenção de contextos, a saber: manuais, revistas científicas, dicionários de áreas conexas, dicionários de língua geral, sítios da Internet e demais fontes que se mostrarem vantajosas para a obtenção de fragmentos de textos que esclareçam o termo ou conjunto de termos que será definido.

Ela é composta essencialmente por cinco tipos de informação: o termo, os contextos definitórios e explicativos, fontes a partir das quais os contextos foram extraídos, data de inserção do contexto na base e nome do membro da equipe que fez a inserção do contexto.

As vantagens principais da base definicional são: a) propiciar um repositório de contextos úteis à compreensão do termo, mesmo que para isso seja necessário agregar contextos definitórios ou explicativos em outras línguas, definições existentes oriundas de obras lexicográficas e terminológicas, ou ainda um contexto de algum termo correlacionado que, ao compreendê-lo, ajudará no entendimento do termo que está sendo investigado; b) identificar as principais informações relevantes do termo que devem constar do corpo das definições coletadas. Por meio da leitura dos excertos compilados, é possível analisar quais são as informações que coincidem e quais conflitam, e o que deve ser considerado, levando-se em conta a procedência (fonte) e o grau de

---

[27] *A kind of symbiosis must develop between these two kinds of specialists. The terminologist's work attests to the linguistics validity of the data; the domain expert's knowledge guarantees that the research stays in line with subject-field reality.* (DUBUC e LAURISTON, 1997, p.87).

[28] Termo cunhado no âmbito do GETerm.

[29] Na seção 5, contextos definitório e explicativo são abordados mais detalhadamente.

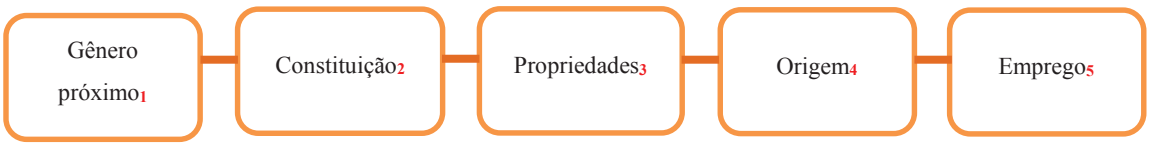
relacionamento do contexto com a área do conhecimento que está sendo descrita; c) possibilitar a recuperação ágil dos contextos nos quais a definição se baseou. Essa recuperação é válida tanto para a revisão da definição, como para sua validação por parte do especialista de domínio; d) observar a evolução do significado do termo ou conjunto de termos ao longo do tempo, o que é útil para se compreender como a área do conhecimento está se desenvolvendo e qual está sendo o comportamento da sua terminologia.

Dessa forma, a base definicional é um recurso adicional à composição do *corpus*, configurando-se como um verdadeiro facilitador da redação da definição terminológica.

3.3.4 Template

Um último recurso mencionado quanto à redação da definição terminológica diz respeito à elaboração de uma espécie de *template*, modelo que serve de orientação sobre quais informações semânticas devem ser inseridas no corpo da definição e qual é a melhor ordem de apresentação. O *template* é criado a partir da leitura dos contextos definitórios e explicativos inseridos na base definicional, referentes a um conjunto de termos que pertencem ao mesmo campo semântico. A título de ilustração, apresentamos no Esquema 3 o *template* utilizado no campo semântico "matéria-prima" da terminologia de Revestimento Cerâmico e, logo a seguir, a definição terminológica de "carbonato de cálcio", termo que integra o campo das matérias-primas e que, por isso mesmo, adequa-se ao *template*.

Esquema 5 – *Template* do campo semântico "matérias-primas" da terminologia de Revestimento Cerâmico



Fonte: ALMEIDA, SOUZA E PINO (2007)

01	**CARBONATO DE CÁLCIO.** [Substância]$_1$ [sólida branca, insolúvel em água, que se decompõe por aquecimento formando-se óxido de cálcio (cal viva) e dióxido de carbono.]$_2$ [Ocorre na natureza como os minerais calcita e aragonita.]$_4$ [Fundente e branqueador, torna o esmalte mais duro e resistente, além de propiciar baixo coeficiente de expansão.]$_3$ [É a matéria-prima mais utilizada para introduzir cálcio em massas e esmaltes. É empregado na composição da maioria dos esmaltes (fusão: 2.095 a 2.485ºC).]$_5$

Como se pôde observar, não necessariamente a ordem é seguida à risca, muitas vezes, por questão de elegância de redação, inverte-se a posição das informações sugeridas pelo *template*.

Chegou-se a esse *template* devido à observação das informações semânticas recorrentes nos contextos que explicam os termos alocados no campo semântico "matéria-prima". Nos contextos, reconhecem-se as informações semânticas por meio de marcadores linguísticos que indicam gênero próximo: "é um", "é um tipo de"; constituição: "consiste em"; propriedades: "tem como característica"; origem: "se encontra em" e emprego: "pode ser usado em". Tais marcadores e suas respectivas informações semânticas têm sido objeto de pesquisas terminológicas, com o propósito de auxiliar na etapa da redação da definição, como nos trabalhos de Seppällä (2004) e Kamikawachi (2009).

A fim de exibir o funcionamento do *template*, seguem alguns contextos enumerados, em vermelho, a partir dos quais foram consideradas as informações semânticas do termo "carbonato de cálcio" para a redação da sua definição.

02	[Precipitated calcium carbonate, in low micron sizes]2, [is used as an inorganic filler in "basing cements"]5.
03	[These cements consist of a two-stage phenol-formaldehyde resin, calcium carbonate filler an enough hexamethylenetetramine to catalyze the reaction of the resin with heat]2. Sometimes various organic dyes are added.[Material can also be used for insulating coatings for ceramic capacitors and printed circuits]5.
04	[Sal cálcica del ácido carbónico, de fórmula CaCO3]2. [Polvo blanco, insoluble en agua. Por efecto del calor (825 graus Celsius) desprende CO2, transformándose en óxido cálcico (cal viva).]3 [Se encuentra en la naturaleza como aragonito, calcita, piedra caliza, mármol, etc.]4 [Se emplea en construcción, fabricación de cementos y cal viva, y en la industria química.]5

05	(CaCO3): [es la materia prima más usada para introducir calcio em pastas y esmaltes.]5 [Se expende en dos formas: natural y artificial.]3 [La primera proviene del mineral calcita o rocas calizas molidas amalla No 200]2, [y es la preferida para la fabricación de pastas]5 (suele llamarse carbonato de calcio "pesado").
06	[Existe también una forma más pura, obtenida químicamente por precipitación (es el carbonato de calcio "liviano").]2(...)
07	Óxido de Cálcio. [Fundente, insolúvel, refratário e branqueador, torna o esmalte mais duro e resistente além de baixo coeficiente de expansão.]3 [Usado na composição da maioria dos esmaltes.]5 [Fusão:2095 a 2485 graus Celcius.]3
08	[um sólido branco, de fórmula CaCO3, que é pouco solúvel na água.]2 [O carbonato de cálcio decompõe-se por aquecimento formando-se óxido de cálcio (cal viva) e dióxido de carbono.]3 [Ocorre na natureza como os minerais calcita e aragonita. As rochas contendo carbonato de cálcio dissolvem-se lentamente sob a ação de chuvas ácidas (contendo CO2 dissolvido) provocando dureza temporária.]4 No laboratório, o carbonato de cálcio é precipitado borbulhando dióxido de carbono na solução aquosa de cal viva. [O carbonato de cálcio é usado na produção de cal (óxido de cálcio) por aquecimento]5

Além de o recurso ser aplicado nas pesquisas no âmbito do GETerm, ele tem sido empregado em diferentes grupos de pesquisa, tais como Faber (2002) e Alcina e Valero Doménech (2008) com o propósito de dinamizar a etapa da definição, pois ao obter o modelo ou *template* de definição, o terminólogo pode observar e procurar mais atentamente o fragmento textual no qual é fornecida determinada informação semântica.

4. CONTEXTO TERMINOLÓGICO

Nesta seção, tornaremos a abordar a noção de contexto e, mais especificamente, as características do contexto definitório (ou definição natural) no âmbito da Terminologia, tratados mais superficialmente na seção 3.

Ambos os objetos – contexto e contexto definitório – se configuram como foco da nossa pesquisa e por isso são detalhados e discutidos aqui. Ao final da seção, após as considerações feitas acerca do tema, espera-se ter explícita nossa posição teórica, com o propósito de servir de orientação no percurso metodológico de identificação e extração dos contextos definitórios em *corpus* especializado.

Primeiramente, é apresentada a relevância do contexto nos estudos terminológicos (4.1). Em seguida, é exibida a classificação de contextos proposta por terminólogos que se ocuparam do tema (4.2).

4.1 RELEVÂNCIA DOS CONTEXTOS NA TERMINOLOGIA

O contexto, compreendido como o entorno linguístico de um termo em um enunciado (AUGER e ROUSSEAU, 1978, p.34), desempenha importante papel no trabalho terminográfico, pois a partir da sua constituição, é possível observar a relação do termo com o campo de conhecimento no qual ele se situa. O contexto comprova a existência, o uso e o funcionamento de determinado termo e pode ainda fornecer informações semânticas sobre ele.[30]

Tais contextos são também imprescindíveis na a) elaboração da ontologia, já que é de onde se extraem os relacionamentos entre os termos do mesmo domínio; b) na redação da definição terminológica, pois por meio deles se obtêm as informações semânticas sobre os termos, tais como hiperônimo, causa, finalidade, origem, e etc.; c) e ainda servem como exemplos de uso do termo, utilizados na ficha terminológica do projeto ou já no verbete, pois a partir do termo em uso é possível observar seu

[30] Célestin et al. (1990, p.30-31) defendem a importância de distinguir dois grandes tipos de contextos na Terminologia: o microcontexto – ambiente imediato do termo estudado, ou seja a frase ou uma parte da frase onde aparece, e o contexto geral no qual se situa o termo, que pode ser um documento inteiro, um capítulo, uma obra (...), este é denominado macrocontexto.

comportamento morfossintático, geralmente sem o compromisso de apresentar ao usuário contexto com informações semânticas do termo ao qual se refere.

Na subseção 4.2, são apresentados os principais tipos de contextos empregados no âmbito da Terminologia.

4.2 TIPOLOGIA DE CONTEXTOS

Dubuc (1978, p.30-31[31]) considera que os termos podem ocorrer em três categorias distintas de contextos. São eles:

a) **Contexto associativo:** contexto no qual se pode localizar o termo estudado em um determinado campo de aplicação, por associação com os termos que o rodeiam, como nos casos em que ele aparece em uma lista enumerada, ou quando é empregado com um valor puramente funcional;
b) **Contexto explicativo:** contexto no qual se resenha de maneira sumária a natureza ou um dos aspectos do termo;
c) **Contexto definitório:** contexto no qual se oferecem dados com indicações sobre o conceito ao qual corresponde o termo em questão, mesmo que o contexto não apresente a forma de uma definição em sentido estrito.

Embora todos os contextos sejam relevantes na elaboração de produtos terminográficos, os contextos que fornecem mais informação sobre o significado do termo são considerados mais preciosos do que aqueles que têm um caráter mais ilustrativo. Isso porque além de eles desempenharem um papel essencial na compreensão do termo, os contextos de caráter explicativo ou definitório são mais raros de se encontrar.

Considerando a quantidade e qualidade de descrição que o contexto fornece, Dubuc e Lauriston (1997, p.82-83), dezenove anos depois, reclassificam novamente os contextos, atribuindo-lhe características um pouco diferentes da versão de Dubuc (1978):

a) **Contexto associativo**: contexto que não fornece nenhuma informação do conceito coberto pelo termo, mas mostra que o termo é usado na linguagem especializada.

[31] Traduzido e adaptado por nós.

b) **Contexto explicativo**: contexto que cria uma imagem aproximada do conceito coberto pelo termo;
c) **Contexto definitório**: contexto que contém descritores em quantidade e qualidade suficiente para expressar uma imagem muito clara do conceito coberto pelo termo, a partir do qual a definição verdadeira facilmente poderia ser inferida.

Pode-se afirmar que a análise terminológica de um contexto e, por conseguinte, sua categorização, passa essencialmente por três importantes variáveis: o termo (TE), as informações semânticas acerca do termo (IS), e sua área do conhecimento (AC). Dessa forma, um contexto para ser um "contexto terminológico" de fato deve essencialmente: estar relacionado com a área do conhecimento que está sendo descrita, caso contrário, o contexto não terá validade alguma. Além disso, o termo procurado precisa estar explícito, senão não é considerado um contexto seu. Já as informações semânticas sobre o termo, ou a falta delas, revelam se o contexto é associativo, explicativo ou definitório. Seguem alguns exemplos de contextos, supondo o cenário da identificação de contextos do termo "spin" para a elaboração de um glossário da área de Nanociência e Nanotecnologia (N&N):

09	O spin é uma propriedade quântica do elétron, mas pode ser interpretado, classicamente, como se o elétron estivesse em permanente rotação em torno de um eixo, como o planeta Terra faz numa escala muita maior.
10	O diagrama de nível de energia para Co2 (configuração 3d7), em um campo ligante octaédrico e tetraédrico, apresenta três transições de spin permitidas.
11	A Chevrolet acaba de lançar seu mais novo carro: o spin.
12	Este trabalho, iniciado em março de 2006, tem como objetivo a elaboração de um sistema de controle de atitude, voltado para pequenos satélites, utilizando o campo magnético terrestre.
13	(g) Spin.

A nona sentença, extraída de um artigo científico do domínio de N&N, fornece o termo "spin" e sua definição. Devido ao fato de o artigo se configurar como um texto da área-objeto, podemos considerar que o fragmento é de fato um contexto terminológico e é do tipo definitório.

Na décima sentença, o termo "spin" está manifestado no mesmo artigo da sentença anterior, portanto, é um contexto terminológico. Porém, como não apresenta nenhuma informação semântica sobre o termo, é classificado como associativo.

A décima primeira sentença, extraída do site da fabricante de veículos *Chevrolet*, apresenta a expressão "spin" e até há alguma informação semântica sobre ela ("é um carro e é novo"), entretanto, não está associada à área de N&N e tampouco se refere ao mesmo conceito ou significado do termo utilizado no domínio. Logo, embora seja um contexto explicativo, não pode ser considerado como um contexto terminológico válido para a N&N.

Já a décima segunda sentença, extraída de um artigo de N&N, não apresenta o termo "spin" e, por consequência, nenhuma informação semântica sua. Dessa forma, não representa um contexto terminológico para o termo "spin", embora possa ser útil para outros termos como "campo magnético terrestre" ou "satélite".

O último exemplo, no qual "spin" ocorre isoladamente, foi também extraído de um texto da área de N&N. E como se nota, é desprovido de informação semântica sobre o termo. Assim, de acordo com a tipologia oferecida por Dubuc (1978) e Dubuc e Lauriston (1997), entendemos que ele é considerado um contexto terminológico e categorizado como associativo, pois é o "testemunho" (CABRÉ, 1992, p.288-289) de sua existência na área-objeto.

Na ISO 12620 (1999, p.25[32]) que versa sobre *Computer applications in terminology – Data categories,* o tópico que trata de contextos se situa dentro do tópico maior de "descrição relacionada ao conceito", o qual se subdivide em definição, explicação, ilustração, exemplo e contexto. Nessa perspectiva, os contextos podem se dividir em:

a) **Contexto associativo:** contexto que contém uma quantidade mínima de informação conceitual necessária para associar um conceito a um campo do conhecimento em particular.

14	*Machine tool operations such as blanking, piercing, lancing, shearing, beading and flanging can also be performed in a press brake.*

[32] Tradução e adaptação feita por nós.

b) **Contexto explicativo**: contexto que fornece uma explicação sumária de um conceito.

15	*The "reed", which keeps the warp yarns separated, helps to determine cloth width.*

c) **Contexto definitório**: contexto que contém informação substancial sobre um conceito, mas que não possui um rigor na forma de definição. Quando um contexto definitório é composto por um discurso prolongado sobre o conceito, pode-se admitir que se trata de um contexto enciclopédico.

16	*Weaving is a method of producing cloth by interlacing two or more sets of yarns, at least one warp and one filling set, at right angles to each other.*

d) **Contexto linguístico**: contexto que ilustra a função de um termo em discurso, mas não fornece nenhuma informação conceitual. Casos em que a ocorrência do termo não é acompanhada de um discurso circundante são considerados apenas *attestations*.

17	*Cylindrical grinders consume relatively little power.*

e) **Contexto metalinguístico**: contexto que consiste em um discurso sobre o termo como um signo de tal forma que o termo é usado de maneira autônoma. O contexto metalinguístico difere de outros tipos de contextos no sentido de que constitui um discurso sobre o termo em si, enquanto que os outros tipos de contextos constituem um discurso sobre o objeto ou a noção ao qual o termo se refere.

18	*The term expertise in French when it is used to mean compétence d'expert (expert competence) is a borrowing from English.*

A classificação proposta é interessante no sentido de detalhar os contextos nos quais os termos podem se manifestar. Esse modo mais minucioso de se analisar os contextos terminológicos pode ser objeto de pesquisas linguísticas de caráter teórico ou aplicado. Contudo, nesta pesquisa, acreditamos ser contraproducente segregar os contextos em variadas tipologias, uma vez que nosso foco é rastrear as marcas linguísticas que evidenciam um contexto como portador de uma ou mais informações semânticas definitórias sobre dado termo, fazendo com que nossa análise inicial sobre o contexto seja

binário – ou é contexto dito genericamente definitório ou não é. Portanto, mesmo que reconheçamos que os contextos apresentam nuanças específicas e que, possivelmente, contextos explicativos, definitórios e metalinguísticos estejam no rol de contextos genericamente definitórios que serão analisados, não entraremos no mérito nesse nível de descrição.

A próxima seção (5) abrange algumas propriedades do contexto definitório, essenciais na sua compreensão e identificação.

PARTE 2

OBJETO DE ESTUDO

5. CONTEXTO DEFINITÓRIO: ASPECTOS PRÁTICOS

> *'If you wish to converse with me,' said Voltaire, 'define your terms.' How many a debate would have been deflated into a paragraph if the disputants had dared to define their terms! This is the alpha and omega of logic, the heart and soul of it, that every important term in serious discourse shall be subjected to the strictest scrutiny and definition. It is difficult, and ruthlessly tests the mind; but once done it is half of any task.*
>
> **Will Durant**

Seguindo a orientação do filósofo Voltaire, com o intuito de esclarecer a noção de contexto definitório empregada nesta pesquisa, esta seção é destinada à discussão de alguns pontos importantes sobre o tema, a saber: definição em texto (como e por que se faz?) (5.1); o termo e o contexto definitório (5.2); contexto definitório *versus* definição (5.3); contexto definitório *versus* contexto explicativo (5.4); delimitação do contexto definitório (5.5) e, finalmente; proposta de (re)definição de contexto definitório (5.6).

5.1 DEFINIÇÃO EM TEXTO (COMO E POR QUE SE FAZ?)

A definição em texto ou em discurso[33] cumpre o papel de conceituar e familiarizar o indivíduo com um termo, inserido em determinado contexto comunicativo.

Além de o tema ser amplamente debatido do ponto de vista epistemológico, especialmente pela Linguística e Filosofia[34], é também de interesse prático daqueles que de alguma forma dele necessitam: a) especialistas e pesquisadores de distintas áreas, como Medicina, Direito, Psicologia e etc. para se aprofundarem nos seus temas de trabalho; b) elaboradores de dicionários, glossários, ontologias e afins, com o propósito de identificar e descrever relações e significados dos termos do universo investigado; c) e até um usuário comum que digita "apendicite" no buscador *Google*, com a intenção de saber o que é a doença.

Esses três grupos de potenciais interessados em consultar ou produzir uma definição reconhecem um fragmento de texto como uma explicação ou definição sobre

[33] As expressões definição em texto e definição em discurso estão sendo usadas como sinônimas de contexto definitório.

[34] De acordo com o linguista Flowerdew (1992, p. 204), a Filosofia tipicamente foca em instâncias descontextualizadas da definição, enquanto que a Linguística Aplicada lida, ou deveria lidar, com definições como elas ocorrem no contexto.

um dado termo, devido a determinados marcadores linguísticos ou tipográficos que fazem o elo entre o *definiens* (definição) e o *definiendum* (termo).

Geralmente produzimos uma definição quando um termo ou conceito surge pela primeira vez em uma situação comunicativa e ele precisa ser esclarecido para que se minimize a "confusão de ordem conceitual" na interação. Se assumirmos que o nosso interlocutor ou audiência desconhece ou conhece muito superficialmente o termo ou conceito expresso para prosseguir a interação, é necessário que façamos uso de recursos linguísticos de forma a esclarecê-lo, associando um significado a ele. Assim, é de se esperar que, no caso de textos acadêmicos (artigos, relatórios, teses, entre outros), sejam eles escritos ou apresentações orais, a definição seja uma estratégia bem utilizada em seções como resumo, introdução e contextualização da pesquisa, uma vez que são espaços reservados para apresentação do tema de estudo. De acordo com Aluísio (1995):

> Algumas formas de definições (definição por substituição e semi-formal) e as descrições são encontradas em várias partes de uma introdução. Mas a estratégia de familiarização aparece, principalmente, no componente de contextualização quando termos devem se tornar conhecidos. Dependendo do público-alvo do artigo, uma quantidade maior ou menor de conceitos é familiarizada [sic]. (Aluisio, 1995)

Ressalte-se que, em qualquer texto, as escolhas lexicais e os recursos que utilizaremos para definir estão subordinados ao nosso entendimento do termo, à proficiência que temos da língua, ao julgamento que fazemos do nosso interlocutor, enfim, à situação comunicativa em que se dá a interlocução.

Seguem alguns trechos da seção "Introdução" de um artigo científico intitulado *Capacidade de retenção de barreiras de proteção produzidas com solo arenoso estabilizado quimicamente* (RIZZO e LOLLO, 2006), os quais exemplificam como ocorre a estratégia de atribuir significado ao termo em início do texto científico, familiarizando a audiência com o termo em questão:

---

19 **O termo** barreira de proteção **é utilizado para designar** camadas de baixa permeabilidade**,** constituídas de materiais naturais, artificiais ou da combinação de ambos, e que têm como objetivo proteger o meio vizinho da percolação de fluidos, sendo utilizados em diversos tipos de obras como canais, reservatórios, diques, lagoas de rejeito, lagoas de tratamento de resíduos e aterros sanitários (Leite, 1997).

---

20 Os processos físicos **são aqueles responsáveis pela** movimentação dos compostos pelos poros do solo, especialmente a advecção e a dispersão. A advecção **pode ser descrita como** o movimento de translação na direção do fluxo da água subterrânea, no qual o soluto se move por estar ligado à água circulante no meio (Leite, 1996). Na dispersão **considera-se que** a substância toma direções diferentes daquela do fluxo principal, espalhando-se, misturando-se e ocupando um volume de solução maior que aquele que ocuparia caso ocorresse apenas o processo de advecção.

21 Sorção **é o termo que descreve** os processos nos quais os solutos (íons, moléculas e compostos) são repartidos entre a fase líquida e a interface da partícula do solo. A complexação **ocorre quando** um cátion metálico reage com um ânion que funciona como ligante inorgânico. Os íons metálicos que podem ser complexados por ligantes inorgânicos incluem metais de transição e metais alcalinos terrosos.

22 **Os tipos de** estabilização **mais comuns são** a mecânica **(**na qual a mistura de um ou mais materiais, antes da compactação, tem por objetivo aumentar a resistência mecânica**)** e a estabilização química **(**na qual se adiciona ao solo uma substância que aumente a sua coesão ou que reduza sua permeabilidade**).**

Nos quatro exemplos elencados, estão sublinhados os termos que estão acompanhados de uma definição ou explicação. É de se notar que o artigo apresenta uma série de termos e definições, muito provavelmente com o intuito de propiciar ao leitor certo grau de conhecimento sobre os principais conceitos que se relacionam com o objeto de estudo, antes de apresentar a problemática e discussão do tema em si.

Ainda que o leitor não conheça o tema, ao analisar a superfície textual, é possível assinalar determinadas expressões linguísticas (em negrito), chamadas de padrões definitórios, que indicam que os fragmentos constituem uma definição ou explicação.

Por outro lado, parte-se do princípio de que as unidades terminológicas usadas na definição (como no ex. 19, o termo "camadas de baixa permeabilidade") são conhecidas pela audiência, caso contrário compromete-se a eficácia do texto definitório. Obviamente que, dependendo do domínio, da estrutura da definição e se a audiência for heterogênea, o uso de termos mais opacos torna-se difícil.

Nesse sentido, alguns linguistas aplicados trabalharam na compreensão dos mecanismos da língua que estão envolvidos nesse tipo de atividade – a de produção de definições em textos autênticos, geralmente com o foco em como formular definições em textos científicos no ensino de inglês para não nativos (PEARSON, 1998, p. 89).

Esse tipo de exploração científica foi realizado por autores como Swales (1971, 1981), Widdowson (1979, 1985), Selinker, Trimble e Trimble (1976), Darien (1981) e Trimble (1985). Os respectivos trabalhos são sumariamente apresentados nas próximas subseções.

5.1.1 Swales

Swales teve como motivação ensinar alunos não falantes de inglês como formular definições nesta língua em contexto acadêmico. De acordo com o autor, definições sempre começam com artigo indefinido, pois são declarações genéricas. Entretanto, conforme aponta Pearson (1998, p.90), citação com a qual concordamos, "não é verdadeiro concluir que todas as definições começam com artigo indefinido" (tradução nossa), como é o caso de termos encabeçados por *uncountable names* (ex. *aluminum, cement*), em que artigos indefinidos não são aceitáveis, como é caso da sentença usada pelo próprio autor:

23 *Aluminum is a metal produced from bauxite* (Swales, 1971, p.70).

Em língua portuguesa, uma definição pode apresentar o termo, antecedido do artigo definido, indefinido ou sem artigo, como nas sentenças 24, 25 e 26 que foram extraídas do nosso *corpus* de estudo para exemplificar o fenômeno.

24 O CAA **é um** equipamento eletromecânico com bombeamento pulsátil do sangue pulsátil construído por duas câmaras: direita e esquerda.

25 Uma rede neural artificial (RNA) **é uma** rede de numerosos elementos computacionais não lineares, altamente interconectados, operando em paralelo e organizados em padrão semelhante a uma rede neural biológica (Lippmann, 1987), e que adquirem conhecimento através da experiência.

26	Reciclean **é um** produto comercial próprio para limpeza de sistemas de irrigação e é fabricado pela empresa Kemira (...)

Além disso, Swales sugere que o verbo mais comum usado para definir é o *is*, e que o sintagma *can be defined* às vezes é utilizado. Ele propõe a seguinte fórmula de definição:

An {x,y} *is a/an* general class word + *wh-* word...
onde *x* é um nome contável e *y* é um nome incontável.

O autor ainda afirma que a definição pode ser acrescentada com oração na voz ativa, como em:

27	*A dentist is a person who takes care of people's teeth* (SWALES, 1971, p.68).

Ou na voz passiva:

28	*A knife is an instrument which is used for cutting things* (SWALES, 1971, p.69).

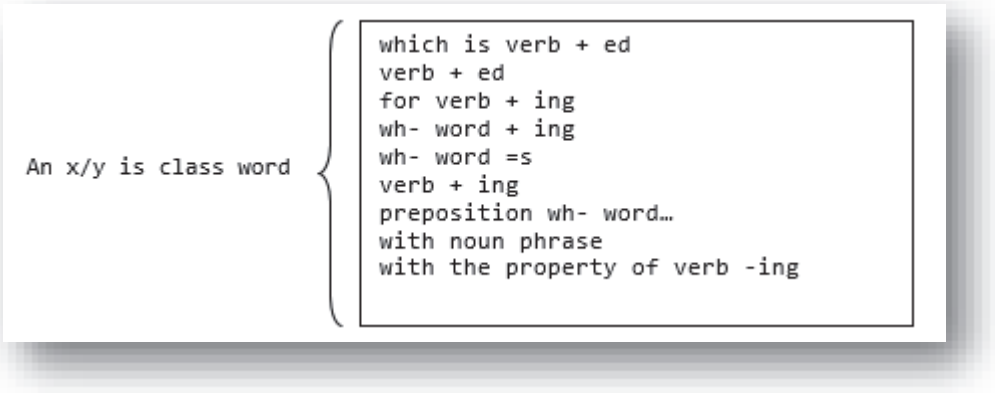
Da mesma forma, a definição pode ser acrescentada com oração reduzida, como em:

29	*Aluminium ia a metal produced from bauxite* (SWALES, 1971, p.70)
30	*A tangent is a straight line touching a curve a tone point* (SWALES, 1971, p.72).

Em ambas as sentenças o *produced* e o *touching* são empregados no lugar do *wh-* da fórmula de definição proposta por Swales. Outras formas do verbo *to be* (p. ex. no passado e futuro) não são comuns em sentenças definitórias como verbo principal, do mesmo modo que é incomum o uso de outros verbos como tal (SWALES,1971, p.68).

Swales reúne as fórmulas de definição no Esquema 4 (SWALES, 1971, p.74).

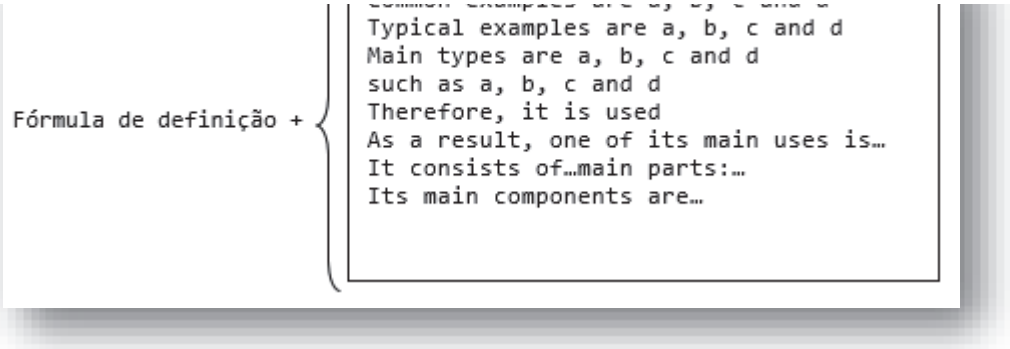
Esquema 6 – Fórmulas de definição propostas por Swales (1971, p.74)



Fonte - SWALES (1971, p.74)

Contudo, o pesquisador observa que as definições não são circunscritas necessariamente apenas a uma sentença. Elas podem se expandir em duas sentenças ou mais, seguindo as fórmulas do Esquema 5.

Esquema 7 – Fórmulas de definição que ocorrem em mais de uma sentença, propostas por Swales (1971, p.74)



Fonte - SWALES (1971, p.74)

Uma constatação feita por Swales (1981, p. 107), e que pôde ser comprovada na nossa pesquisa, é que definições são raras em artigos de pesquisa (porém comuns em textos didáticos). A razão disso, na visão do autor, é que a função das definições é mais fornecer explicações de termos que estabelecer axiomas que formam parte de um sistema lógico de postulados e teoremas (SWALES, 1981).

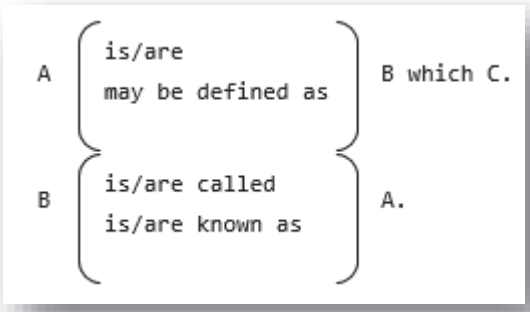
Embora seu trabalho levante questões interessantes acerca da formulação de definições em texto, lamentavelmente Swales usa frases construídas do cotidiano para exemplificar suas afirmações sobre definições em comunicação acadêmica. No nosso

ponto de vista, ainda que essas definições construídas sirvam para compreendermos as fórmulas propostas, elas enfraquecem os argumentos do autor, já que ele não utiliza dados autênticos.

5.1.2 Widdowson

A investigação de formulação de definições em textos por Widdowson teve como contexto o ensino de inglês para estudantes que precisam conhecer a linguagem com o objetivo de prosseguir seus estudos em ciências e tecnologia na educação superior (Widdowson, 1979). Para atingir essa meta, Widdowson, em coautoria com Allen, produziram no âmbito da série *English in Focus*, a obra didática *English in Physical Science*, que oferece exercícios de definição. O autor cita, por exemplo, duas formas comuns de definição em discurso científico (1985, p.81), como se pode observar no Esquema 6.

Esquema 8 – Fórmulas de definição em discurso científico (WIDDOWSON, 1985, p.81)



Fonte – WIDDOWSON (1985, p.81)

Da mesma forma que Swales, Widdowson também opta por não utilizar exemplos extraídos de textos autênticos em seus exercícios, tentando, assim, simplificar a formulação de definições.

5.1.3 Selinker, Trimble e Trimble

Os autores assumem que a definição é a função básica para o pensamento científico em relatórios (SELINKER, TRIMBLE e TRIMBLE, 1976). Por meio dela, o leitor tem acesso às seguintes informações: o termo que é definido, a classe da qual o termo é membro e a declaração de características essenciais ou diferenças que distinguem o termo de outro termo da classe. Ao contrário dos autores citados em 5.1.1 e 5.1.2, Selinker *et al.* afirmam que a definição pode estar esparsa no texto, dentro de um ou mais parágrafos, cujos principais objetivos retóricos são "descrição", "explicação", "classificação" ou "apresentação de informações sobre os procedimentos experimentais".

Os autores deixam explícito que o método de pesquisa teve como base o uso de textos autênticos. Provavelmente por esse motivo, eles conseguiram olhar um pouco mais para o entorno das definições e observar com mais atenção o fenômeno linguístico. Além de outras coisas, o uso de um *corpus* permitiu chegarem à conclusão de que muitos exemplos fornecidos em exercícios de livros didáticos de *English for Science and Technology* (EST) não existem em textos autênticos.

5.1.4 Darien

Darien afirma que a definição é um "acordo" sobre termos, o qual fornece a base para nossa comunicação, seja ela cotidiana ou especializada. Por isso, a definição de palavras e conceitos é uma das nossas mais importantes competências analíticas. Para o autor:

> A definição é mais bem compreendida como um conjunto de sistemas encaixados, dominados pelo sistema semântico, que interage com a subordinação sintática e com os sistemas lexicais e tipográficos para produzir uma ampla variedade de fórmulas de definição. (DARIEN, 1981, p. 43)[35] (tradução nossa).

É possível observar que o autor reconhece que a definição pode se manifestar de variadas formas e que, além do componente semântico, o léxico, a sintaxe e elementos tipográficos também estão envolvidos nessa atividade. A título de exemplo, uma definição por paráfrase (componente semântico) pode ser expressa por uma frase apositiva (componente sintático) e entre vírgulas, como em: "Pterodáctilos, répteis

[35] *Defining is a best understood as a series of interlocking system dominated by the semantics system, which interacts with the subordinate syntactic, lexical and typographic system to produce a broad range of definition formulas.* (DARIEN, 1981, p. 43)

semelhantes a pássaros com penas escamosas, viveram no Período Jurássico Inferior" (tradução nossa) (DARIEN,1981, p.48).[36] Assim, se fosse usado outro componente semântico, a definição possivelmente teria outras nuanças léxicas, sintáticas e tipográficas.

Darien também apresenta uma estrutura geral da definição formal, composta pelo termo, um gênero próximo ou uma palavra genérica e uma ou mais diferenças específicas e delimitadas. Cada um desses elementos da definição é explicado e exemplificado em seu trabalho.

5.1.5 Trimble

Trimble também teve como escopo o ensino de definição em inglês para não nativos. Ele distingue definições em simples e complexas. A definição simples é expressa em uma única sentença, e a definição complexa é manifestada em duas ou mais sentenças. Além disso, o autor propõe três tipos de definições simples: a definição formal, a semiformal e a não-formal.

A definição formal fornece ao leitor três tipos de informação: o nome do termo que será definido, a classe à qual o termo pertence e a(s) diferença(s) entre o termo e todos os outros membros da classe, na seguinte estrutura, que nada mais é o paradigma GPDE, abordado na subseção 3.1 deste trabalho:

A genus is a species which + distinguishing characteristic.

A informação que se apresenta nesse tipo de definição se refere à descrição física, função, uso e objetivo. A seguinte sentença "Um anemômetro é um instrumento meteorológico que registra a velocidade do vento em uma tela ou escala"[37] (tradução nossa) (TRIMBLE, 1985, p.75-76) é um exemplo fornecido de uma definição formal que apresenta a descrição da sua função. Ainda que ele forneça exemplos de definição formal, de acordo com Pearson (1998, p.98), o autor não indica quão comum é esse padrão de definição e se há outras formas de expressar essas informações na definição.

---

[36] *Pterodactyls, birdlike reptiles with scaly feathers, lived in the Lower Jurassic Period.* (DARIAN,1981, p.48)

[37] *An anemometer is a meteorological instrument that registers the speed of wind on a dial or gage.* (TRIMBLE, 1985, p.80)

A definição semiformal contém apenas dois dos três elementos definitórios básicos: o termo definido e a declaração das diferenças (TRIMBLE, 1985, p.77). O autor sugere que a ausência do gênero próximo na definição se deve ao fato de o escritor assumir que essa informação é óbvia ou irrelevante para a discussão, como em "Um anemômetro registra a velocidade do vento em uma tela ou escala"[38] (tradução nossa) (TRIMBLE, 1985, p. 77). Porém, Trimble não faz menção sobre quais palavras podem ser usadas como elo entre o termo e as características distintivas (PEARSON, 1998, p.99).

Por último, a definição não-formal fornece ao leitor dois tipos de informação: o termo definido e outra palavra ou frase que tem o significado aproximado do termo ou que ainda apresenta uma característica marcante do termo, como por exemplo: "Um aracnídeo é uma aranha"[39] (tradução nossa) (TRIMBLE, 1985, p.80).

As pesquisas citadas apresentam pontos em comum quanto ao foco no ensino de inglês para não nativos e no contexto acadêmico. Além disso, com exceção do trabalho de Selinker, Trimble e Trimble (1976), os demais fazem uso de exemplos construídos como forma de demonstrar seus argumentos. Isso nos faz entender que suas pesquisas não foram baseadas em amostras autênticas de textos, e que, portanto, podem não passar de intuição linguística. Não que as fórmulas de definição não sejam realizadas na língua, mas o fato de não se ter a análise das ocorrências reais de uso dá margem para uma suposição ingênua ou de se ter em mãos um fenômeno que não é produtivo na língua. De qualquer maneira, as características apontadas na formulação de definições em texto nessas pesquisas servem como "sementes" para serem lançadas em *corpus* fértil, as quais podem germinar e dar bons frutos.

Após essa brevíssima incursão na Linguística Aplicada, em busca do que é definição e de como esta se formula, as próximas subseções destinam-se a prestar esclarecimento sobre o que estamos considerando como contexto definitório a partir do que foi relatado sobre os tipos de contextos na Terminologia e também do que foi posto sobre definição em texto pelos linguistas aplicados.

5.2 BUSCA DE CONTEXTO DEFINITÓRIO POR TERMO

[38] *An anemometer registers the speed of wind on a dial or gage.* (TRIMBLE, 1985, p. 80)
[39] *An arachnid is a spider.* (TRIMBLE, 1985, p.80)

Voltando para o cenário da Terminologia, o lugar natural para analisar os termos para a construção da ontologia (ou estrutura conceitual) e para a redação da definição terminológica é o próprio texto produzido pela ou para a comunidade usuária da terminologia em questão. Como já mencionado na subseção 4.2, nos textos, os termos podem aparecer em: a) contextos em que eles somente são mencionados, sem constar qualquer informação conceitual sobre o termo – contexto associativo; b) contextos que criam uma imagem aproximada do conceito coberto pelo termo – contexto explicativo e; c) contextos que contêm descritores em quantidade e qualidade suficiente para expressar uma imagem muito clara do conceito coberto pelo termo – contexto definitório (DUBUC e LAURISTON, 1997, p.82).

Ao submetermos o mesmo artigo científico usado no início da subseção 5.1, agora completo, a um concordanciador[40] qualquer, pudemos observar que o termo "barreira(s) de proteção" ocorre outras vezes ao longo do texto:

31	Capacidade de retenção de **barreiras de proteção** produzidas com solo arenoso estabilizado quimicamente
32	O artigo avalia a capacidade de retenção de **barreiras de proteção** ambiental de misturas de solo-cimento (7% em massa) e solo cal (8% em massa) para substâncias de origem orgânica (chorume e água residuária) e inorgânica orgânica e valores promissores para as soluções inorgânicas.
33	Palavras-chave: **Barreiras de proteção**, capacidade de retenção, solo-cal, solo-cimento, estabilização de solos.
34	O termo **barreira de proteçã**o é utilizado para designar camadas de baixa permeabilidade, constituídas de materiais naturais, artificiais ou da combinação de ambos, e que têm como objetivo proteger o meio vizinho da percolação de fluidos, sendo utilizados em diversos tipos de obras como canais, reservatórios, diques, lagoas de rejeito, lagoas de tratamento de resíduos e aterros sanitários (Leite, 1997).
35	Boff (1998) afirma que diversos autores indicam que para atender ao uso como **barreira de proteção** o material deve apresentar como propriedades básicas valores de condutividade hidráulica inferiores a 10-9m/s e eficiência na retenção das substâncias potencialmente contaminantes de interesse.

[40] Concordanciador é uma ferramenta computacional utilizada em processamento de *corpus* para listar as ocorrências de uma determinada expressão linguística que pode conter uma ou mais palavras, a qual fica em destaque, com uma quantidade definida de palavras à sua direita e à sua esquerda.

36	A condutividade hidráulica obtida com as misturas nas condições de compactação utilizadas indica valores compatíveis para uso como liners, viabilizando o uso de solos lateríticos arenosos para produção de **barreiras de proteção.**

Como é possível notar, os fragmentos de número 31, 32, 33 e 36 são exemplos de contexto associativo, pois o termo se manifesta na sentença, porém não é apresentada nenhuma informação conceitual/semântica sobre ele. Já o fragmento 35 disponibiliza informações semânticas de "propriedade" ou "caracterização" da "barreira de proteção", podendo, portanto, ser considerado como um contexto explicativo. E por último, o fragmento 34 é o que se apresenta mais completo e se caracteriza como um contexto definitório, pois apresenta uma série de informações semânticas que o descrevem, tais como "o que é", "do que é constituído", "sua finalidade" e "onde é aplicado". Dessa forma, se considerarmos o trabalho da redação terminológica desse termo, possivelmente apenas os fragmentos 34 e 35 seriam escolhidos para serem consultados, ou seja, apenas eles constariam da base definicional[41] do termo "barreira de proteção".

5.3 CONTEXTO DEFINITÓRIO *VERSUS* CONTEXTO EXPLICATIVO

Todavia, ainda que se reconheça que existam contextos que são mais completos, pois apresentam mais informações semânticas do termo (como o exemplo 34) que outros (como o exemplo 35), devido ao fato de o contexto definitório muitas vezes se confundir com o contexto explicativo e devido ao fato de os contextos explicativos também serem muito úteis no trabalho da redação terminológica, nesse trabalho não será feita distinção entre eles. Aqui, ambos serão considerados contextos definitórios, tendo, portanto, o adjetivo "definitório" o sentido de contexto auxiliar na redação da definição.

Assim, ambos os fragmentos, 34 e 35, são considerados contextos definitórios.

5.4 CONTEXTO DEFINITÓRIO *VERSUS* DEFINIÇÃO

Nas publicações sobre o tema, em geral, os termos "contexto definitório" e "definição" concorrem. Neste trabalho, será privilegiada a expressão "contexto

[41] Ver subseção 3.3.3.

definitório" em detrimento da expressão "definição", pelas seguintes razões: a) o termo foi incorporado em parte considerável das pesquisas em Terminologia[42] no Brasil (ALMEIDA, PINO e SOUZA, 2007; BARROS, 2004) e em outros países (SAGER, 2003; CASTILLO, 1997 e DUBUC, 1999), b) pela contraposição ao termo "contexto associativo"; c) para não gerar ambiguidade com o termo "definição" usado no sentido de "definição dicionarizada"[43]; d) pelo fato de o termo "contexto definitório" compreender mais nitidamente, além da definição ou explicação, o termo e o elemento de ligação entre eles.

Portanto, deste ponto em diante, passamos a chamar "definição em texto" de "contexto definitório", ainda que os autores citados utilizem a expressão "definição".

5.5 DELIMITAÇÃO DO CONTEXTO DEFINITÓRIO

Em termos práticos, os contextos definitórios podem abranger um fragmento entre parênteses, uma oração, uma sentença, um parágrafo ou mais. Nesse sentido, podemos relembrar Trimble (1985), que distingue os contextos definitórios em simples, quando expressos dentro de uma única sentença, e complexos, quando são expressos em mais de uma sentença.

Na verdade, o que se percebe é que é uma tarefa árdua quando temos de delimitar no texto o que é o contexto definitório. Talvez isso seja mais simples, quando, por exemplo, o contexto se apresenta como uma classificação ou uma oração adjetiva, como ocorre no exemplo 22 da subseção 5.1.

Em outros momentos, é mais difícil precisar qual é a linha limítrofe do contexto definitório, quando, por exemplo, é feita uma exposição sobre o tema, como no exemplo 19, exibido na subseção 5.1.

Esta dificuldade existe, entre outros fatores, porque a quantidade e a profundidade de informações imprescindíveis no contexto definitório para a redação da definição terminológica dependem do projeto que está sendo desenvolvido. Pode ser que para um dicionário destinado a iniciantes, a consulta a uma ou duas linhas seja o

[42] <www.btb.termiumplus.gc.ca/tpv2alpha/alpha-por.html?lang=por&i=1&index=ptt&srchtxt=contexto%20definitorio>. Acesso em ago. 2012.

[43] <www.btb.termiumplus.gc.ca/tpv2alpha/alpha-por.html?lang=por&i=1&index=ptt&__index=ptt&srchtxt=defini%E7%E3o&comencsrch.x=-425&comencsrch.y=-288&comencsrch=Iniciar>. Acesso em ago. 2012.

suficiente, enquanto que para outros projetos, voltados para um público mais especializado, seja necessária uma quantidade maior de informações.

5.6 DEFINIÇÃO DE CONTEXTO DEFINITÓRIO

A respeito das considerações feitas sobre contexto definitório nesta seção da tese, propomos uma definição de "contexto definitório" do ponto de vista da Terminologia, a qual guiará o nosso percurso de pesquisa:

Contexto definitório: fragmento textual que tem como função familiarizar um termo e seu significado a uma dada audiência. Está contido em uma ou mais sentenças, apresenta o termo ou expressão que remete a ele (anáfora) e uma ou mais informações semânticas ligadas ao termo por meio de uma expressão linguística (é definido como, tem a função de, e etc.) ou elemento tipográfico (dois pontos, parênteses, entre outros). É utilizado como suporte na construção de ontologias e na redação da definição terminológica. **Sinônimo:** Definição; Definição em texto.

Agora que já foi realizada uma breve discussão sobre o tema e algumas de suas nuanças, na seção 6 será apresentada a revisão bibliográfica acerca de trabalhos que têm como foco a tarefa de reconhecimento e extração de contextos definitórios.

6. IDENTIFICAÇÃO E EXTRAÇÃO DE CONTEXTOS DEFINITÓRIOS

> *'Odio las definiciones' es un dicho atribuido a Benjamin Disraeli. Si su propósito fue señalar la extrema dificultad de definir las palabras, uno tendría que estar de acuerdo con él (...) De hecho, todo el problema de la definición es más interesante de lo que muchos suponen. (...) podemos aprender mucho sobre la naturaleza del significado de la palabra examinando la teoría y la práctica de la definición.*
>
> **John Lyons**

Nesta seção 6, como sugere o linguista John Lyons, será feita uma incursão na "prática da definição", a fim de observar como o nosso objeto de estudo é constituído e, mais do que isso, conhecer de que forma ele tem sido analisado e empregado nas diferentes pesquisas. Para tal, nos valemos de alguns pontos de vista de estudiosos que vêm investigando o tema a partir de diferentes propósitos, mas que têm em comum a necessidade de, ao final do trabalho, ter dados formalizados ou formalizáveis que sirvam de substrato linguístico em sistemas de processamento de língua natural (PLN) que apresentam como recurso a identificação e extração de contextos definitórios.

Dessa forma, a finalidade desta seção é apresentar o estado da arte quanto a investigações realizadas sobre a formalização dos contextos definitórios. Inicialmente, com o intuito de esclarecer o percurso dessa etapa de revisão bibliográfica, é descrita, de modo sucinto, a sequência adotada desde a busca pelos textos que constituiriam parte da revisão. Em seguida, são apresentados trabalhos que se ocuparam, em alguma medida, de sistematizar o conhecimento linguístico referente aos contextos definitórios. Cada pesquisa é exposta individualmente, de modo que serão salientados aspectos como a) descrição geral e língua de aplicação; b) constructo; c) *corpus* e metodologia; d) resultados e avaliação e e) comentários.

6.1 PERCURSO DESTA REVISÃO BIBIOGRÁFICA

A revisão bibliográfica tem como propósito apresentar o panorama no qual a pesquisa está inserida, como também sustentar o trabalho e revelar alguma lacuna em relação ao tema de pesquisa. Assim, a revisão bibliográfica é considerada como o passo inicial para qualquer pesquisa científica (WEBSTER e WATSON, 2002).

A revisão bibliográfica, de caráter sistemático, foi utilizada nessa parte do trabalho, sobretudo, porque havia a necessidade de recuperar os trabalhos feitos nesse domínio, de modo que o presente estudo estivesse em consonância com o que já foi feito, com o intento de preencher vazios e dar prosseguimento a encaminhamentos sugeridos pelos pesquisadores em seus trabalhos. Além do fato de que, nesta seção 6, está o tema fulcral desta tese.

De forma geral, nesse percurso seguimos o roteiro proposto por Conforto, Amaral e Silva (2011), os quais organizam a revisão bibliográfica sistemática em três fases – entrada, processamento e saída. A seguir, cada uma dessas fases é descrita.

6.1.1 Entrada

Grosso modo, essa etapa inicial consiste na **definição** do problema de pesquisa e dos **objetivos** que orientam a realização da revisão bibliográfica sistemática. Isso feito, são selecionadas algumas **fontes primárias** de estudo a partir de critérios pré-estabelecidos e, a partir dessa eleição, são construídas as ***strings* de busca** e são definidos os **critérios de inclusão e qualificação** dos textos encontrados. Por fim, são selecionados o **método** e as **ferramentas** que serão utilizadas nas buscas dos textos, no armazenamento e na manipulação dos arquivos, além da proposta de um cronograma a ser executado.

Neste estudo, buscamos responder à seguinte questão: como recuperar contextos definitórios em *corpus* escrito? Esse **problema** se caracteriza como ponto de partida da pesquisa. Alinhado com o problema que se deseja investigar, a revisão bibliográfica tem como **objetivo** principal recuperar trabalhos que apresentam alguma sistematicidade linguística quanto à identificação e à recuperação de contextos definitórios. A fim de alcançar esse objetivo, foram consultadas nossas **fontes primárias**, as quais foram identificadas de acordo com as seguintes circunstâncias: a) conhecimento prévio ao tema (AUGER, 1997); b) sugestões recomendadas por especialistas da área (PEARSON, 1998) e (WENDT, 2010) e; c) base de dados do Instituto Universitário de Linguística Aplicada (IULA)[44] (ALARCÓN, 2009). Esse conjunto de trabalhos permitiu reconhecer termos, autores e textos relevantes do domínio em questão.

[44] Instituto pertencente à Universidade Pompeu Fabra, Barcelona, Espanha (<www.iula.upf.edu/>, acesso em ago. 2012).

Na etapa da formação das ***strings* de busca**, graças ao estudo preliminar das fontes primárias citadas anteriormente, foi possível identificar e sistematizar alguns termos relacionados ao tema da pesquisa, os quais seriam utilizados nas buscas em bases referenciais de artigos e *abstracts* como *Web of Science*[45] e *Scopus*.[46] Da mesma forma, outros sistemas de busca de trabalhos científicos seriam consultados, como *Google Acadêmico*.[47] Vale ressaltar que as bases possuem mecanismos distintos quanto à construção de *strings* e de uso dos operadores lógicos, o que exigiu uma consulta ao manual ou ao tópico de ajuda do ambiente.

A princípio, foram estabelecidas as *strings* exibidas no Quadro 1. O primeiro e o terceiro blocos referem-se a *strings* nomeadas "satélite", pois seu conjunto diz respeito a expressões que podem acompanhar o núcleo, ou seja, são facultativas e foram estipuladas com o intuito de otimizar as buscas. Já as *strings* nomeadas "núcleo" podem ser empregadas autonomamente ou acompanhadas dos satélites nas buscas nas bases referenciais e representam a parte principal da expressão de busca.

45 <apps.webofknowledge.com>. Acesso em ago. 2012.
46 <www.scopus.com/home.url>. Acesso em ago. 2012.
47 <scholar.google.com.br>. Acesso em ago. 2012.

Quadro 1 – *Strings* de busca.

	Português	Inglês	Espanhol	Francês
Satélite	extração	extraction	extracción	extraction
	identificação	identification	identificación	identification
	exploração	exploitation	exploración	exploration
	recuperação	retrieval	recuperación	récupération
	descrição	description	descripción	description
	análise	analysis	análisis	analyse
Núcleo	contexto definitório	definitional context defining context definitory context	contexto definitorio	contexte définitoire
	definição	definition	definición	définition
	enunciado definitório	-	-	énoncé définitoire
		term definition	-	-
	padrão definitório	-	-	-
Satélite	corpus	corpus	corpus	corpus
	texto	text	texto	texte

Fonte - Elaborado pela autora.

Essa etapa de delimitação das expressões de busca foi essencial na pesquisa, haja vista que "definição", como se trata de uma expressão geral, é amplamente empregada em títulos, resumos, palavras-chave de textos acadêmicos e científicos de diversas áreas, tendo como consequência um retorno imenso de textos ruídos.[48] Além disso, foi relevante fazer um levantamento dos termos sinônimos de "definição" de modo a resgatar textos que apresentam essa mesma temática.

Outro ponto de destaque é a equivalência das expressões em: a) inglês, que se constitui como a língua na qual a divulgação da ciência é mais intensa; b) português, pois é a língua de descrição deste trabalho e; c) espanhol e francês, devido ao fato de ambas línguas gozarem de certa tradição nos estudos terminológicos e na área de PLN também.

Em seguida, foram elaborados os **critérios de inclusão** dos textos que seriam selecionados para a pesquisa a partir das *strings*. Esses critérios são úteis na medida em que podem auxiliar no descarte (e na inclusão) de textos, de acordo com alguns princípios

[48] A título de exemplo, em 29/08/13, no sítio da *Web of Science*, a expressão "definition" no campo "Title" retornou mais de 50 mil publicações. Conferir em: <apps.webofknowledge.com/summary.do?SID=4Bb57ZZcmTBWV ne7lFJ&product=UA&qid=1&search_mode=GeneralSearch>. Acesso em ago. 2013.

norteadores da pesquisa, assim como em relação a alguns quesitos externos à pesquisa. No nosso escopo, os seguintes critérios foram instituídos:

a) Quanto ao foco da pesquisa [CI-1]: trabalhos que descrevem algum nível de sistematização na identificação e extração de contextos definitórios;
b) Quanto à área [CI-2]: trabalhos provenientes da Computação (Processamento de Linguagem Natural, Mineração de textos e Recuperação e Extração de Informação), Linguística (Terminologia, Lexicologia, Lexicografia) e Ciência da Informação;
c) Quanto ao tipo textual [CI-3]: trabalhos redigidos no formato de artigo, tese, dissertação, capítulo de livro, relatório técnico e registro de patente.

Após a seleção de textos por meio das *strings* de busca e dos critérios de inclusão, os textos escolhidos foram submetidos ao conjunto de **critérios de qualificação**, o qual é especialmente útil para avaliar a importância do documento para o estudo. Esse item serviu para agrupar textos por similaridade e para ordená-los em ordem de prioridade na nossa revisão. Os critérios foram baseados (ordem aleatória):

a) Na popularidade [CQ-1]: textos mais citados devem ser priorizados, pois podem significar que são trabalhos significativos.
b) Na data [CQ-2]: quanto mais recente for a publicação do texto, maior é a chance de se obterem trabalhos que representam o *status quo* da área.
c) Na origem [CQ-3]: textos oriundos de grupos de pesquisa que investigam definições devem ser analisados antes daqueles que têm um caráter mais disperso, em razão da possibilidade de recuperar informações mais elaboradas nos trabalhos inseridos em grupos de pesquisa.
d) Na língua [CQ-4]: textos referentes à descrição de contextos definitórios em português, uma vez que se trata da língua de trabalho da presente pesquisa.

O passo seguinte foi realizar a escolha do **método** e das **ferramentas.** Para a condução inicial das buscas foram previstas as bases: *Scopus*,[49] *Scielo*,[50] *Web of Science*,[51]

[49] <www.scopus.com/>. Acesso em ago. 2013.
[50] <www.scielo.org/>. Acesso em ago. 2013.
[51] <wokinfo.com/>. Acesso em ago. 2013.

Portal Capes de Periódicos,[52] *Google Patent*,[53] *Google Acadêmico*[54] *e Biblioteca Digital Brasileira de Teses e Dissertações*.[55]

Essas foram as bases de conhecimentos inicialmente previstas, contudo, para uma compreensão mais apurada do assunto, após a consulta e as leituras iniciais, caso se fizesse pertinente, seriam sondadas informações sobre o nosso objeto de pesquisa em outros sítios também. As **ferramentas** adotadas para armazenamento e gerenciamento dos textos foram *Mendeley*[56] e o programa de planilha eletrônica *Excel*. *Mendeley* é um *software* gratuito que permite gerir, partilhar, anotar, referenciar e citar artigos científicos e textos diversos.[57] Já no *Excel* é possível organizar e analisar dados extraídos dos trabalhos lidos por meio de tabelas, gráficos e outros recursos, apoiando a sistematização da revisão bibliográfica.

Finalizando essa etapa de planejamento, foi previsto o período de 24 meses para a realização da revisão bibliográfica, sendo que, após o término do período, se algum trabalho fosse publicado e estivesse em consonância com o tema proposto, o mesmo poderia seria incorporado à revisão.

A seguir, será descrita a etapa referente ao processamento da revisão bibliográfica.

6.1.2 Processamento

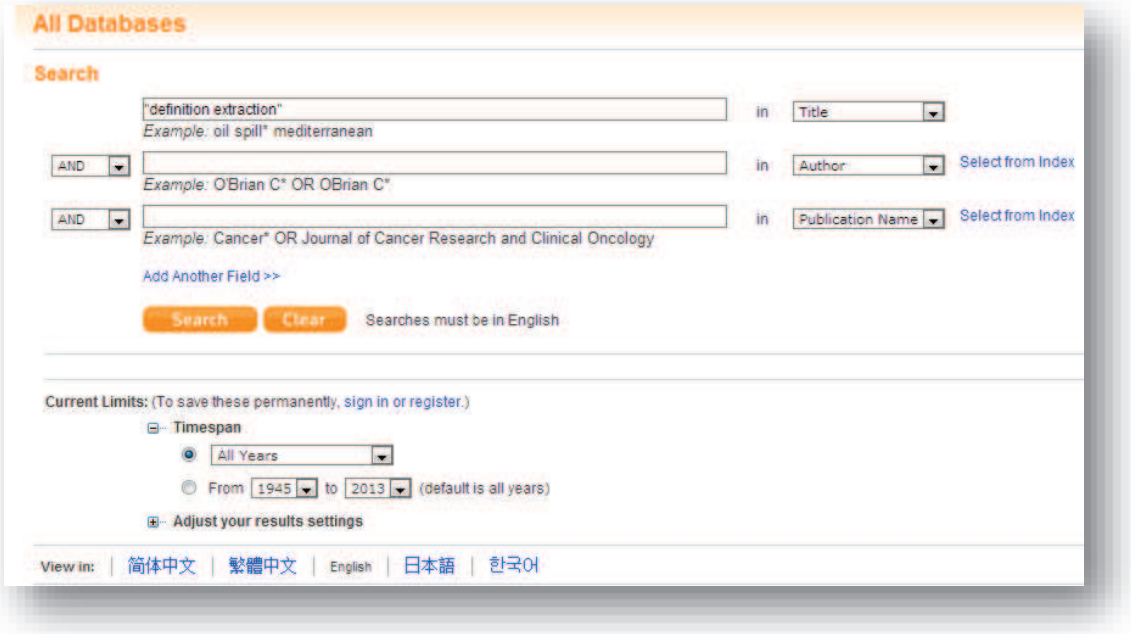
Essa segunda etapa reúne, de fato, a **condução das buscas** nas bases de conhecimento mencionadas anteriormente, a aplicação dos critérios de inclusão e qualificação, **análise dos resultados** obtidos a partir da leitura preliminar dos documentos eleitos e, por fim, a **documentação** que engloba os detalhes dos textos incorporados à pesquisa.

Ressalta-se que, sobretudo, quanto à **condução das buscas**, o método foi iterativo, ou seja, foram feitos refinamento das buscas e criação de novos filtros quando necessário. As Figuras 1 e 2 exemplificam essa etapa.

---

[52] <www-periodicos-capes-gov-br.ez31.periodicos.capes.gov.br/>. Acesso em ago. 2013.
[53] <www.google.com/?tbm=pts>. Acesso em ago. 2013.
[54] <scholar.google.com.br/>. Acesso em ago. 2013.
[55] <bdtd.ibict.br/>. Acesso em ago. 2013.
[56] <www.mendeley.com/>. Acesso em ago. 2013.
[57] <bibliotecafea.com/tag/mendeley/>. Acesso em ago. 2012.

Figura 1 – Página de busca da base *Web of Science*.

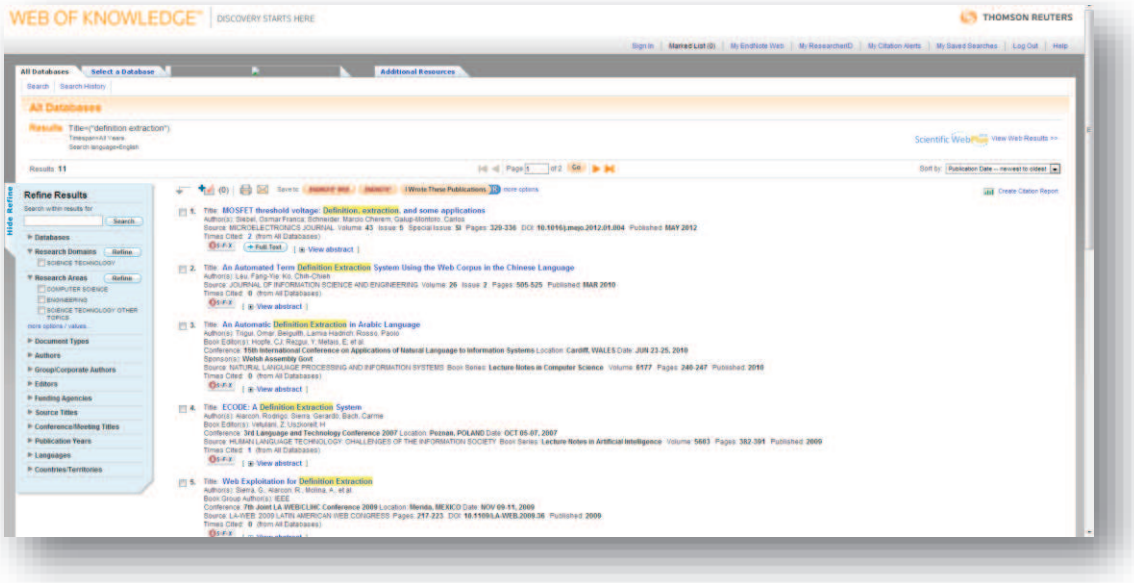


Fonte - *Print screen* da página inicial da base *Web of Science*.

A Figura 1 se refere à página de busca da base *Web of Science* (*Web of Knowledge*), a qual apresenta formulários de texto que podem ser preenchidos por uma ou mais *strings*. O usuário decide sobre qual ou quais campos do texto serão incididas as buscas. Ainda nessa página, é possível limitar o período no qual a busca será feita. Como percebemos, nesse exemplo, a *string* "definition extraction" foi dada como entrada, no campo "Title", considerando todos os textos, independentemente da data de publicação.

Na Figura 2, é exibida a tela com os resultados alcançados a partir da busca feita no formulário da Figura 1. No total, foram recuperados 11 artigos que possuem o sintagma "definition extraction" no título.

Figura 2 – Página com os resultados gerados da base *Web of Science*



Fonte – *Print screen* da página de resultados da base *Web of Science*.

Em sequência, são descritas algumas características observadas acerca dos textos encontrados e que estão em consonância com a pesquisa. É o que se chama, neste percurso, de **análise dos resultados**.

Ao fim de todo o processo, foram encontrados mais de trinta trabalhos sobre o tema, para várias línguas e sob diferentes perspectivas. Segue, no Quadro 2, parte da relação de textos encontrados que abordam a identificação e extração de contexto definitório.

Quadro 2 – Relação de trabalhos sobre contexto definitório.

TÍTULO	AUTOR(ES)	LÍNGUA	ANO
ECODE: A Definition Extraction System	Alarcón, R., Sierra, G., & Bach, C.	ESPANHOL	2009
Developing a Definitional Knowledge Extraction System	Alarcón, R., Sierra, G., & Bach, C.	ESPANHOL	2009
Repérage des énoncés d ' intérêt définitoire dans les bases de données textuelles	Auger, A.	FRANCÊS	1997
Automatic Definition Extraction Using Evolutionary Algorithms	Borg, C.	INGLÊS	2009
Definition Characterisation through Genetic Algorithms	Borg, C., Rosner, M., & Pace, G. J.	INGLÊS	2008
Automatic Grammar Rule Extraction and Ranking for Definitions	Borg, C., Rosner, M., & Pace, G. J.	INGLÊS	2007
Towards Automatic Extraction of Definitions	Borg, C., Rosner, M., & Pace, G. J.	INGLÊS	2007
Supporting e-learning with automatic glossary extraction: Experiments with Portuguese	Del Gaudio, R., & Branco, A.	PORTUGUÊS	2007
Learning to Identify Definitions using Syntactic Features	Fahmi, I., & Bouma, G.	INGLÊS	2005
Grammar-based Automatic Extraction of Definitions	Iftene, A., & Trandabăţ, D.	ROMENO	
Repérage et exploitation d ' énoncés définitoires en corpus pour l ' aide à la construction d ' ontologie	Malaisé, V., Zweigenbaum, P., & Bachimont, B.	FRANCÊS	2004
Mining defining contexts to help structuring differential ontologies	Malaisé, V., Zweigenbaum, P., & Bachimont, B.	FRANCÊS	2005
Extracção de Definições no Corpógrafo	Pinto, A. S.	PORTUGUÊS	2004
Mining on-line sources for definition knowledge	Saggion, H., & Gaizauskas, R.	INGLÊS	2004
Hacia un sistema de extracción de definiciones en textos jurídicos	Sánchez,A., Márquez,M.	ESPANHOL	2005
Análisis, diseño e implementación de un agente deliberativo para extraer contextos definitorios en textos especializados	Suaréz de La Torre, M. M., & Al, E.	ESPANHOL	2009
Automatic extraction of definitions from German court decisions	Walter, S., & Pinkal, M.	ALEMÃO	2006
Geração automática de glossários de termos específicos de um corpus de Geologia	Wendt, I. da S.	PORTUGUÊS	2010
Definition Extraction using Linguistic and Structural Features	Westerhout, E.	HOLANDÊS	2009
Ranking definitions with supervised learning methods	Xu, J., Cao, Y., Li, H., & Zhao, M.	INGLÊS	2005

Fonte – Elaborado pela autora.

Observamos que, entre as áreas mencionadas na seção anterior, a Computação é a que mais apresenta publicações sobre o tema, com 12 dos 20 trabalhos apontados no Quadro 2.

Tal interesse dos informatas pelo tema se deve à busca pela melhoria constante dos sistemas de perguntas e respostas, sistemas de buscas, enfim, pela recuperação da informação de modo mais otimizado.

Ademais, no Quadro 3 são exibidos dois projetos ambiciosos que se destacam quanto ao tema.

Quadro 3 – Projetos sobre contexto definitório.

Grupo	Instituição	Língua
Language Technology for e-Learning(LT4eL)	Universidade de Utrecht e parceiros	búlgaro, checo, holandês, inglês, alemão, maltês, polonês, português europeu e romeno
Grupo de Ingeniería Lingüística(GIL)	Universidade Nacional Autónoma de México	espanhol

Fonte – Elaborado pela autora.

Ambos os empreendimentos são apresentados logo à frente, porém, vale ressaltar que são projetos nos quais a identificação e extração de contexto definitório fizeram-se presentes em uma série de trabalhos, os quais foram elaborados por pesquisadores distintos, com o intuito de subsidiar sistemas de recuperação de definição, a partir da descrição e formalização desse tipo de conhecimento linguístico.

Na pesquisa bibliográfica realizada, constatamos ainda, especificamente, dois eventos internacionais importantes sobre o tema, explicitados no Quadro 4.

Quadro 4 – Eventos realizados sobre contexto definitório.

Evento	Promoção	Ano	Publicação
1st International Workshop on Definition Extraction	UNAM	2009	www.aclweb.org/anthology/W/W09/W09-4400.pdf
Text REtrieval Conference – TREC: Question and Answer	NIST	2003	trec.nist.gov/pubs/trec12/papers/OVERVIEW.12.pdf

Fonte – Elaborado pela autora.

Realizado em conjunto com o *Recent Advances in Natural Language Processing (RANLP),* a proposta do *1st International Workshop on Definition Extraction* foi justamente promover um espaço de discussão de cunho teórico e aplicado acerca da extração de definição no âmbito do PLN. Como primeiro evento, o *workshop* veio preencher uma lacuna no cenário acadêmico e científico, o qual carecia de um espaço de compartilhamento de informações sobre metodologias, ferramentas, técnicas de avaliação ou aplicações relacionadas a essa temática (SIERRA, POZZI e TORRES, 2009, p. iii).

Já o *workshop Text Retrieval Conference* (TREC),[58] em 2003, teve como foco avaliar sistemas baseados em *Question Answering* que, a partir de uma grande coleção de documentos, recuperam respostas classificadas como: factoides ("What city is the home to the Rock and Roll Hall of Fame?"; listas ("*What grapes are used in making wine*?") e definições ("*Who was Abraham in the Old Testament?*" ou "*What is Ph in biology?*")[59]. Como afirmam Saggion e Gaizauskas (2004, p.1), a recuperação automática de respostas

[58] <en.wikipedia.org/wiki/Text_Retrieval_Conference>; <trec.nist.gov/data/qa/t2003_qadata.html>. Acesso em ago. 2012.

[59] Para mais exemplos, consulte <trec.nist.gov/data/qa/2003_qadata/03QA.tasks/test.set.t12.txt>. Acesso em ago. 2012.

do tipo definicional em fontes não estruturadas são de grande importância, tanto para consultas *ad hoc* quanto para a construção automática de glossário.

Em relação à língua, foram encontradas duas iniciativas significativas sobre o tema para o português europeu: Pinto e Oliveira (2004), no contexto do Corpógrafo, que será explicado adiante, e Del Gaudio e Branco (2007), no contexto do *Language Technology for e-Learning* (LT4eL); e uma iniciativa para o português brasileiro (WENDT, 2010), oriunda do grupo de pesquisa em PLN da Pontifícia Universidade Católica do Rio Grande do Sul (PUC-RS).[60]

Na etapa de **documentação,** todos os textos encontrados e que têm potencial valor para a pesquisa foram catalogados e armazenados no *Mendeley*, o qual já foi citado anteriormente na etapa "entrada". Além do *software*, os textos também foram armazenados em uma planilha *Excel* para composição de tabelas comparativas, a partir de dados extraídos dos trabalhos.

A terceira e última etapa da revisão bibliográfica é descrita na subseção 6.1.3.

6.1.3 Saída

Esta última etapa da revisão bibliográfica consistiu no registro dos textos que foram finalmente agregados à pesquisa. Em seguida, foi realizada uma síntese da bibliografia estudada, a qual é exposta na subseção 6.2.

De acordo com Conforto, Amaral e Silva (2011, p.10), a importância dessa etapa no processo científico reside no fato de que "é essencial para identificar o estado atual do corpo de conhecimento no assunto pesquisado, por exemplo, descrevendo os principais autores da área, a evolução do conceito, quantidade de artigos diretamente relacionados ao tema de pesquisa (...)".

Por esses motivos apontados, houve o esforço de sistematizar um procedimento de recuperação e gerenciamento dos textos relacionados ao tema da pesquisa, para que tivéssemos condições de visualizar o panorama do objeto de estudo do modo mais límpido possível, não com a intenção de esgotar o tema ou ainda de resolver de uma vez por todas a descrição e extração de contextos definitórios, mas sim com o propósito de proporcionar dados suficientes para esta pesquisa, assim como para contribuir com outros estudos que estão por vir.

[60] <www.inf.pucrs.br/~linatural/>. Acesso em fev. 2014.

Com essa etapa cumprida, os trabalhos mais relevantes, segundo nossa perspectiva, são resenhados adiante.

6.2 PRINCIPAIS TRABALHOS SOBRE EXTRAÇÃO DE CONTEXTOS DEFINITÓRIOS

Na presente subseção, são abordados os trabalhos realizados sobre extração de contextos definitórios que julgamos serem mais relevantes, de acordo com os critérios determinados no roteiro da revisão bibliográfica. Por uma questão de organização, a apresentação da revisão será realizada particularmente para cada trabalho (ou grupo de trabalho ou ainda produto), com destaque para os seguintes aspectos: a) descrição geral e língua de aplicação; b) constructo; c) *corpus* e metodologia; d) resultados e avaliação; e por fim, e) comentários.[61]

6.2.1 Corpógrafo

DESCRIÇÃO GERAL E LÍNGUA DE APLICAÇÃO

Desenvolvido a partir de 2003 pela Linguateca (polo CLUP/FLUP[62]), o Córpografo[63] consiste em uma plataforma *web* e *desktop* livre para as línguas portuguesa, espanhola, inglesa, italiana, francesa e alemã, que permite colecionar textos em vários formatos, formar e analisar *corpora*, extrair terminologia e criar bases de dados terminológicas com a possibilidade de codificar relações semânticas e ontologias (MAIA; SARMENTO; SANTOS, 2005). Para tal, o Corpógrafo disponibiliza uma série de funcionalidades integradas em uma interface amigável e que pode ser utilizada para fins de investigação em Linguística, Tradução, Engenharia do Conhecimento, além da própria Terminologia.

Um dos recursos que a plataforma fornece a favor da Terminologia é um extrator semiautomático de definições. Como salientam Pinto e Oliveira (2004), o trabalho foi um dos primeiros a explorar o tema, não tendo à época, portanto, muitas informações disponíveis na literatura em relação às fases do processo de extração de definições.

---

61 Esse modo de revisão bibliográfica se baseia no modo de apresentação do trabalho de Alarcón (2009).
62 Centro Linguística da Universidade do Porto/Faculdade de Letras da Universidade do Porto.
63 <www.linguateca.pt/corpografo/>. Acesso em ago. 2012.

CONSTRUCTO

Nesse estudo, consideramos como contexto definitório, "toda a frase ou segmento de frase que possa descrever/caracterizar um dado termo" (PINTO e OLIVEIRA, 2004, p.2).

CORPUS E METODOLOGIA

O *corpus* de estudo de 54.340 mil palavras, nomeado de "Neurodemo", compreendeu textos científicos sobre "neurônios" em português europeu e inglês. Após o pré-processamento do *corpus*, foi realizada a extração de termos do domínio para que, a partir desses termos, fossem observadas e geradas as estruturas sintáticas que fazem com que uma frase ou segmento de frase, que contém o termo, seja considerado uma definição. Para o português europeu, foram extraídos 270 termos; e para o inglês, 661. Após a identificação dessas estruturas, elas foram representadas por meio de expressões regulares na linguagem *Perl*, tendo seu processo de construção sido efetuado de forma iterativa. Em seguida, todos os padrões construídos também foram aplicados a 56 textos em português europeu (439.215 palavras) e 107 em inglês (441.990 palavras) das áreas de PLN e Linguística, o que gerou a validação dos padrões existentes, como também um aumento de padrões encontrados (PINTO e OLIVEIRA, 2004).

Por meio dos exemplos fornecidos em Pinto e Oliveira (2004), os padrões detectados são apresentados no Quadro 5.[64]

[64] De acordo com Sarmento et al. (2006), no Corpógrafo foram disponibilizados cerca de 135 padrões para o português, além disso, é prevista a possibilidade de o usuário agregar manualmente seus próprios padrões definitórios em suas buscas (ALARCÓN, 2009). Infelizmente, até o fechamento deste texto não foi possível acessar o sistema para conhecer e testar o extrator de definições do Corpógrafo.

Quadro 5 – Padrões definitórios extraídos de Pinto e Oliveira (2004).

(^|^o|^a|^um|^uma|o|a|um|uma)TERMO é (o|a|um|uma)[65]
(^|^os|^as|os|as)TERMO são[66]
(^|^o|^a|o|a)TERMO que é (o|a|um|uma)[67]
(^|^o|^a|^um|^uma|o|a|um|uma)TERMO é (constituíd|determinad|compost)(o|a) (por|pelo|pela|pelos|pelas|de)
(^|^os|^as|os|as)TERMO(originam-se|designam-se|ligam-se|dão-se|estendem-se)

Fonte – Pinto e Oliveira (2004).

RESULTADOS E AVALIAÇÃO

Como forma de avaliação, os padrões definitórios detectados foram submetidos a um *corpus* de Fibromialgia (campo de estudo da Medicina), pois as pesquisadoras pretendiam saber se de fato os padrões encontrados também eram produtivos em outros domínios do conhecimento (PINTO e OLIVEIRA, 2004). O *corpus* de avaliação foi formado por 10 textos em português europeu (21.667 palavras) e 23 textos em inglês (80.295 palavras), e foram selecionados 30 termos do campo de estudo em cada uma das línguas. A Tabela 1 apresenta uma amostra dos resultados para o inglês.

[65] Exemplo: "O neurónio é uma (…)." (PINTO e OLIVEIRA, 2004, p.3).
[66] Exemplo: "Os neurónios são (…)." (PINTO e OLIVEIRA, 2004, p.3).
[67] Exemplo: "(…) núcleo celular, que é a central de energia da célula." (PINTO e OLIVEIRA, 2004, p. 4).

Tabela 1 - Parte dos resultados da avaliação do extrator semiautomático de definições do Corpógrafo

Termos	Definições existentes	Definições corretamente extraídas	Definições extraídas
Autonomicnervous system	2	1	1
Central nervous system	1	0	0
Chronic fatigue syndrome	1	0	0
Chronicpain	1	0	0
Craniosacral system	1	0	0
Craniosacraltherapy	1	1	1
fibromyalgia	31	29	33
Fibromyalgiadiagnosis	1	0	0
Fibromyalgiapain	2	1	1
Fibromyalgias syndrome	3	3	4
...	...	...	...
Total			
30	184	104	130

Fonte – Pinto e Oliveira (2004), adaptado pela autora.

A Tabela 1 exibe alguns termos e seus respectivos resultados quanto ao número de definições existentes no *corpus* (segunda coluna), quanto ao número de definições que foram corretamente extraídas pelo sistema (terceira coluna) e, por último, o número de definições extraídas pelo Corpógrafo (quarta coluna). Ao totalizar os valores de cada coluna da Tabela 1, apresentados integralmente em Pinto e Oliveira (2004), chegamos aos valores inseridos na última linha da tabela.

Com esses resultados, podemos avaliar que os padrões definitórios testados tiveram uma taxa global de acerto em torno de 56% (ou seja, recuperaram-se 104 definições de um total de 184 definições existentes no *corpus*); e de todos os 130 candidatos à definição que o sistema recuperou, 20% não são definições de fato.

COMENTÁRIOS

A proposta de fornecer conhecimento linguístico para um extrator de candidatos a contextos definitórios feita pela equipe do Corpógrafo vai ao encontro do nosso plano, pois além do foco na definição, o contexto no qual ambas as pesquisas se inserem é a mesma, ou seja, a construção de recursos de PLN para a redação da definição terminológica. Ainda que Pinto e Oliveira (2004) tenham descrito muito sumariamente a metodologia e os resultados obtidos com os padrões definitórios elaborados, e embora os

números obtidos na avaliação sejam tímidos para afirmar a validade do extrator, a pesquisa foi pioneira nesse tipo de estudo com foco na terminologia de língua portuguesa.

6.2.2 LT4EL

DESCRIÇÃO GERAL E LÍNGUA DE APLICAÇÃO

No contexto do *e-Learning*, o *Language Technology for eLearning* (*LT4eL*[68]) foi um projeto financiado pela Comunidade Europeia entre 2005 e 2008, que teve como principal objetivo aprimorar o *LMS* (*Learning Management System*),[69] de modo a facilitar a recuperação e gestão de materiais de aprendizagem e de informação (DEL GAUDIO e BRANCO, 2009) em várias línguas: búlgaro, checo, holandês, inglês, alemão, maltês, polonês, português europeu e romeno. Além disso, Monachesi, Lemnitzer e Simov (2006, p.667) afirmam que um dos propósitos do projeto é mostrar que a Tecnologia da Linguagem pode prover solução para a tarefa de desenvolvimento de glossários.[70]

Como consta do projeto mestre, no *Documented glossary candidate detector and integration report* (DOCUMENTED, 2005),[71] os glossários podem ser entendidos como um pequeno recurso lexical que apoia o leitor na decodificação do texto e na compreensão de conceitos centrais que são veiculados. E este glossário pode ser construído a partir dos contextos definitórios que estão presentes nos próprios objetos de aprendizagem.

É justamente nesse escopo do projeto que foram criadas frentes de trabalho nas línguas do consórcio quanto ao desenvolvimento de método para a extração automática de definições, com o objetivo final de proporcionar uma ferramenta que auxiliasse na construção de glossários úteis para autores de objetos de aprendizagem que desejassem compilar glossários, bem como para usuários de objetos de aprendizagem que procurassem pela definição de um termo desconhecido ou dúbio presente nos próprios objetos de aprendizagem.

CONSTRUCTO

---

68 <www.lt4el.eu/index.php?content=about>. Acesso em ago. 2012.

69 A forma em português é "Sistema de Gestão da Aprendizagem (SGA)" e designa *softwares* desenvolvidos sobre uma metodologia pedagógica para auxiliar a promoção de ensino e aprendizagem virtual ou semipresencial. Extraído de: <pt.wikipedia.org/wiki/Learning_management_system>. Acesso em ago. 2012.

70 *Glossary candidate detector*

71 Disponível em: <www.lt4el.eu/extern/files/D2.3.pdf>. Acesso em ago. 2013.

A abordagem na identificação e extração de contextos definitórios do LT4eL segue três eixos importantes para todas as línguas: é baseada em regras gramaticais, em linguagem especializada e em documentos anotados linguisticamente (DOCUMENTED, 2005).

CORPUS E METODOLOGIA

Para cada língua, foi formado um *corpus* de estudo a partir de documentos escritos como tutoriais, teses e artigos. Os textos compilados eram provenientes do domínio de *e-Learning* ou área correlata.

No Quadro 6, é descrito o tamanho de alguns dos *corpora* de estudo usados no projeto para as distintas línguas.

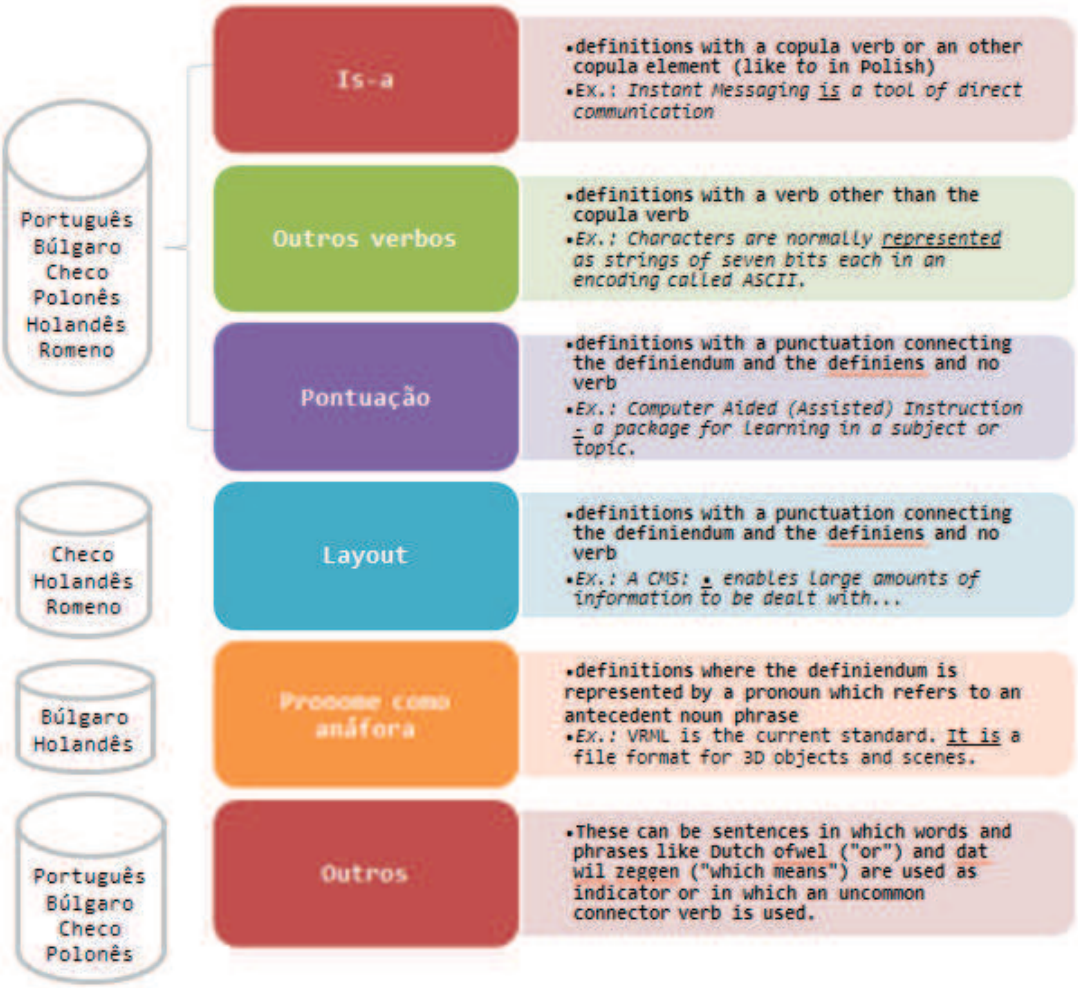
Quadro 6 – Tamanho dos *corpora* de estudo do LT4eL.

Línguas	**Tamanho**
Portuguesa	274.000 mil *tokens*
Romena	700.00 mil tokens
Eslavas (búlgaro, checo e polonês)	Cerca de 83.000 mil *tokens*/língua
Holandesa	Cerca de 505.736 mil tokens

Fonte – DEL GAUDIO e BRANCO (2009), IFTENE, TRANDAB e PISTOL (2007), adaptado pela autora.

A partir da construção de cada um dos *corpora*, foi possível trabalhar na identificação e anotação manual em formato *xml* de contextos definitórios e, por conseguinte, elaborar regras gramaticais para sua extração. Evidentemente, para cada língua foi construída uma gramática local específica, contudo, a tipologia de definições utilizada foi bem aproximada. Além disso, os seis padrões de contextos definitórios são explicitados (em negrito) e exemplificados (em itálico) pelos organizadores do projeto *LT4eL* (DOCUMENTED, 2005), como é possível observar na Figura 3.

Figura 3 – Tipologia de padrões de contextos definitórios do LT4El.



Fonte – DOCUMENTED (2005). Adaptado pela autora.

Observa-se que os três primeiros padrões, "*is-a*", "outros verbos" e "pontuação" foram analisados nas seis línguas do projeto, enquanto que o padrão "pronome como anáfora" foi descrito apenas pelo búlgaro e holandês. Ressalta-se também que, dos seis padrões propostos no projeto, todas as línguas adotaram entre 4 a 5 deles.

RESULTADOS E AVALIAÇÃO

Na etapa de avaliação, os contextos definitórios anotados manualmente nos *corpora* de estudo foram confrontados com os contextos definitórios extraídos

automaticamente em *corpora* de teste, a partir das regras gramaticais constituídas. Com esse fim, foram utilizadas métricas de precisão, cobertura e de F2[72].

Destaca-se que, na nossa apresentação dos dados, foram considerados os resultados somente com *corpora* de teste, com base em correspondência de sentença (*sentence matching*) e não em correspondência de palavras (*token matching*). Além disso, descartou-se a avaliação do F1 da língua holandesa, e por questão de padronização, o resultado apresenta-se na forma decimal (onde 1 equivale a 100%).

Os resultados para o português europeu são apresentados para os três domínios representados no *corpus: Information Society (IS), Information Technology (IT) e e-Learning (e-L)*. Os dois primeiros domínios são compostos por tutoriais, enquanto que o último é composto por artigos e teses (Tabela 2).

Tabela 2 – Avaliação global da língua portuguesa no *LT4eL*.

Corpus	Precisão	Cobertura	F2
IS	0.14	0.86	0.32
IT	0.33	0.69	0.51
e-L	0.11	0.59	0.24

Fonte – DEL GAUDIO e BRANCO (2009). Adaptado pela autora.

Podemos observar que os números obtidos quanto ao *recall*, ou seja, cobertura, foram satisfatórios, enquanto que os valores obtidos quanto à precisão, não. Isso pode ter ocorrido devido ao fato de ter sido oferecido como *input* gramáticas muito simplificadas do verbo "ser" e uma análise superficial dos verbos. Além do mais, poucos exemplares de padrões de definições por pontuação foram encontrados no *corpus* de estudo, sem contar os erros na anotação morfossintática do próprio *corpus* (DEL GAUDIO e BRANCO, 2009).

Na avaliação das línguas eslavas (Tabela 3), nota-se que a precisão foi praticamente a mesma para os três idiomas (0.22). A cobertura do checo e do polonês foi de 0.46 e 0.32, respectivamente, enquanto que a do búlgaro foi de apenas 0.08. Segundo os autores, esses resultados não satisfatórios se devem a uma complexidade das línguas, o que faz com que seja imprescindível a geração de mais regras (PRZEPIÓRKOWSKI et al., 2007).

[72] Medida derivada da F1, que privilegia o desempenho da cobertura em detrimento ao da precisão.

Outra possibilidade levantada pelos pesquisadores quanto à baixa precisão diz respeito às definições formadas por mais de uma sentença (*multi-sentence definitions*). Os autores explicam que, devido ao tipo de anotação adotada (DTD), as definições multissentenças representam um problema, pois a definição só pode ser considerada um subelemento da sentença, e não o contrário. Em casos de definições que ocupam mais de uma sentença, para cada sentença há um elemento separado encapsulando a parte da definição contida na sentença (PRZEPIÓRKOWSKI et al., 2007).

Tabela 3 – Avaliação das línguas eslavas no *LT4eL*.

Língua	Precisão	Cobertura	F2
Búlgaro	0.22	0.08	0.11
Checo	0.22	0.46	0.33
Polonês	0.23	0.32	0.28

Fonte – PRZEPIÓRKOWSKI et al. (2007). Adaptado pela autora.

Quanto ao romeno, foram avaliados os quatro tipos definitórios, e podemos notar que a cobertura foi de 100%, ou seja, o sistema recuperou todos os contextos definitórios do *corpus* que foram extraídos automaticamente por meio das regras propostas para essa língua, como vermos na Tabela 4. Já em termos de precisão, o tipo definitório que se saiu melhor foi o de `verb_def`, seguido do `is_def`. Segundo o autor, os padrões com `is_def` são mais difíceis de serem identificados, pois é um verbo muito profícuo na língua, mas em muitos casos não são constituintes de definição. Quanto aos dois últimos tipos, apesar das várias regras elaboradas, elas não foram consideradas suficientemente precisas (IFTENE, TRANDAB e PISTOL, 2007).

Tabela 4 – Avaliação da língua romena no *LT4eL*.

Tipo definitório	Precisão	Cobertura	F2
is_def	0.53	1	0.77
verb_def	0.75	1	0.90
punct_def	0.14	1	0.33
layout_verb	0.04	1	0.13

Fonte – IFTENE, TRANDAB e PISTOL (2007). Adaptado pela autora.

E por fim, é exibido o desempenho para o holandês (Tabela 5). Percebemos que o sistema apresentou uma cobertura satisfatória somente para os tipos `is_def` e `punct_def`, enquanto que a precisão foi menos de 0.25 em todos os tipos.

Tabela 5 – Avaliação da língua holandesa no *LT4eL*.

Tipo definitório	Precisão	Cobertura	F2
is_def	0.20	0.91	0.43
verb_def	0.25	0.41	0.34
punct_def	0.02	0.76	0.07
layout_verb	0.06	0.40	0.14

Fonte – (Westerhout e Monachesi, 2007). Adaptado pela autora.

COMENTÁRIOS

O projeto descrito teve seu mérito no sentido de contemplar várias línguas para a descrição de um mesmo fenômeno relevante – a definição. Os resultados demonstram que a construção de regras gramaticais para a extração de definições não é tarefa trivial e que é dependente de língua. Ressaltamos também o fato de que no projeto geral o foco foi dado para a cobertura, uma vez que o propósito final era apresentar aos usuários a maior gama possível de contextos definitórios. Dessa forma, há a impressão de que as gramáticas não foram construídas com objetivo de serem altamente precisas.

6.2.3 Definder

DESCRIÇÃO GERAL E LÍNGUA DE APLICAÇÃO

Desenvolvido por Muresan e Klavans (2001), DEFINDER (*Definition Finder*) é um sistema baseado em regras, de identificação e extração de termos juntamente com suas definições em textos orientados a usuários não especialistas de informação médica para a língua inglesa. O sistema foi desenvolvido no âmbito do *Center for Research on Information Access* em parceria com o *Department of Computer Science*, ambos da *Columbia University* e o *output* do sistema pode ser usado em aplicações como: criação e/ou melhoria de recursos terminológicos *on-line*, sumarização e categorização textual de acordo com o nível de especialização – por exemplo, leigos *versus* técnico (KLAVANS; MURESAN, 2001).

CORPUS E METODOLOGIA

Como *corpus*, foram selecionados textos de diferentes gêneros, como artigos de revistas e periódicos, capítulos de livros e manuais escritos por especialistas para leitores

não especializados. Para tanto, as seguintes fontes foram consultadas: *The Merck Manual of Medical Information – Home Edition, Columbia University College of Physician & Surgeons Complete Home Medical Guide; Cardiovascular Institute of the South; Reuters Health Newspaper for Consumers* e *Medical Industry Today.* [73]

O sistema de extração de definições se baseia em dois principais módulos: 1) um de processamento de texto superficial, que busca por determinados padrões léxicos e tipográficos, usando uma gramática de estado finito. Tais padrões são, por exemplo: "é chamado", "é o termo usado para descrever", "é definido como", etc., e mais um conjunto limitado de marcadores textuais, como parênteses e travessão e; 2) um analisador gramatical associado a um analisador sintático estatístico para detectar fenômenos linguísticos mais complexos em contextos definitórios, como aposição e anáfora.

RESULTADOS E AVALIAÇÃO

O sistema foi avaliado qualitativa e quantitativamente por três diferentes vias. A primeira teve por objetivo avaliar o sistema de acordo com as métricas de precisão e cobertura. A segunda avaliou a qualidade do dicionário gerado automaticamente, de acordo com especialistas e não especialistas; e a terceira teve como intento comparar os resultados com dicionários *on-line* da área médica.

Em relação à primeira avaliação, os resultados do DEFINDER foram comparados com resultados manuais da seguinte forma: um *corpus* de teste com nove textos foi construído, seguindo os mesmos critérios utilizados na compilação do *corpus* de estudo. Em seguida, solicitou-se a quatro sujeitos que selecionassem manualmente todas as definições encontradas nesses textos. Determinou-se *gold-standard* quando as definições encontradas eram marcadas por, ao menos, três sujeitos. Os resultados alcançados são os apresentados na Tabela 6.

Tabela 6 - Avaliação automática do *DEFINDER*.

Precisão	Cobertura
86,95%	75,47%

Fonte – KLAVANS e MURESAN (2001).

[73] Não foi encontrada nas publicações sobre o sistema a quantidade de *tokens* ou de textos utilizados no *corpus* de estudo.

Os números indicam que o sistema apresentou na avaliação uma cobertura satisfatória, ou seja, uma grande quantidade de definições foi recuperada (quase 76%). Melhor ainda foi o resultado da precisão, que recuperou 87% de definições corretas, isto é, somente 13% aproximadamente de fragmentos detectados não eram definições de fato.

Na avaliação quanto à qualidade do dicionário gerado automaticamente, sendo julgado por especialistas e não especialistas, a título de comparação, foram utilizados recursos como o *UMLS (Unified Medical Language System)* e o *OMD (Online Medical Dictionary)*. Oito sujeitos não especialistas assinalaram aleatoriamente 15 termos médicos e suas respectivas definições da *UMLS, OMD e DEFINDER*, sem que as fontes fossem reveladas. Os sujeitos deveriam classificar as definições quanto a sua utilidade (utilidade para entender o termo) e clareza (grau de especialização da definição) em uma escala de 1 a 7, em que 1 significava péssimo; e 7, excelente. Os resultados são apresentados na Tabela 7.

Tabela 7 – Avaliação manual do DEFINDER

Usabilidade	**DEFINDER**	5,17
	UMLS	2,94
	OMD	3,9
Claridade	**DEFINDER**	5,65
	UMLS	3,18
	OMD	4,3

Fonte – KLAVANS e MURESAN (2001). Adaptado pela autora.

Os resultados sugerem que o DEFINDER foi mais bem qualificado do que os demais sistemas, tanto em relação à utilidade, quanto à qualidade das definições extraídas, sob o ponto de vista do usuário não especialista.

Para contrastar esses resultados, seguindo a mesma metodologia anterior, 16 sujeitos especialistas classificaram as definições de acordo com a exatidão e integridade, também em uma escala de 1 a 7. Como resultado, o DEFINDER obteve a média de 5,87 para a exatidão e 5,38 para a inteireza, demonstrando que as definições extraídas, além de adequadas para o público leigo, não deixam de ser exatas e completas.

Por fim, os resultados do DEFINDER foram comparados com outros dicionários terminológicos *on-line*. O objetivo desta comparação foi avaliar se o sistema desenvolvido poderia servir como um complemento dos atuais recursos. Foram escolhidos os seguintes dicionários: *UMLS, OMD* e *GPTMT (Glossary of Popular and*

Technical Medical Terms). 93 termos e suas respectivas definições foram extraídos automaticamente pelo DEFINDER e, a partir de três critérios, a avaliação foi feita: 1) o termo estava listado e definido com a mesma definição do DEFINDER em algum dos dicionários; 2) o termo estava listado em algum dos dicionários, porém a definição não era igual à do sistema; 3) o termo não estava na lista em nenhum dos três sistemas.

Tabela 8 – Avaliação manual II do DEFINDER

Critério	UMLS	OMD	GPTMT
1	60% (56)	76% (71)	21.5% (20)
2	24% (22)	-	-
3	16% (15)	24% (22)	78.5% (73)

Fonte – KLAVANS e MURESAN (2001). Adaptado pela autora.

Os resultados dessa avaliação apontam que o DEFINDER apresenta uma quantidade considerável de definições iguais às propostas pela *UMLS e OMD*. Já quanto ao segundo critério, a *OMD* e o *GPTMT* não apresentaram termo que tenha uma definição distinta da proposta pelo DEFINDER, enquanto que a *UMLS* apresenta 24% de definições diferentes das sugeridas pelo sistema. Por último, termos que não constavam das bases da *UMLS* e *OMD* representam cerca de 20%, e quase 79% dos 93 termos buscados na *GPTMT* não foram encontrados.

COMENTÁRIOS

O trabalho de Klavans e Muresan objetivou extrair definições, a partir de técnicas superficiais e profundas de processamento de linguagem natural, de textos médicos orientados para não especialistas. Embora os artigos que apresentam o DEFINDER o fazem de modo muito conciso, não permitindo, portanto, saber, por exemplo, exatamente quantas regras e quantos *tokens* há no *corpus* de estudo, eles ressaltam a avaliação realizada, que, aliás, é uma contribuição interessante para a área, pois vai além da precisão e cobertura, demonstrando que há outras variáveis a serem consideradas nessa tarefa.

6.2.4 Ranking definitions with supervised learning methods

DESCRIÇÃO GERAL E LÍNGUA DE APLICAÇÃO

O artigo *Ranking Definitions with Supervised Learning Methods* é o desdobramento de um trabalho feito por Jun Xu e Min Zhao durante um período em que visitaram a *Microsoft Research Asia*, sob supervisão de Yunbo Cao e Hang Li. Os autores partem do problema de *"definition search"*, isto é, dado um termo de consulta, extraem-se os candidatos a definições do termo (sentenças ou parágrafos) oriundos de vários documentos e, em seguida, é feito o ranqueamento desses candidatos de acordo com o grau de "definitude" (XU et al., 2005).

Assim, o objetivo do trabalho é apresentar uma metodologia para a classificação automática desses fragmentos, de acordo com a probabilidade de serem definições ou não. Os candidatos seriam classificados entre bons, ruins e indiferentes. No trabalho, utilizou-se aprendizado de máquina supervisionado, tendo o algoritmo SVM[74] como modelo de classificação e o Ranking SVM[75] como modelo de regressão ordinal como formalização do problema. Os resultados foram implementados no *Information Desk*, que é um sistema disponível na *intranet* de uma companhia de Tecnologia da Informação, que fornece quatro tipos de busca, sendo o de definição uma das opções.

CONSTRUCTO

Na especificação do projeto, é proposta a classificação dos candidatos em definições boas, definições ruins e definições indiferentes. Contudo, foi necessário criar critérios objetivos para essa distinção.

Entende-se por boa definição, nesse contexto, fragmentos que contenham uma noção geral do termo, que pode vir acompanhada de expressões como *"is a kind of"* e com algumas propriedades do termo. As sentenças de 37 a 39 são exemplos de boas definições.

Uma definição ruim não fornece uma noção geral do termo e nem suas propriedades. São fragmentos que podem ser opiniões, impressões ou sentimentos das pessoas sobre o termo, não sendo possível capturar o significado do termo pela leitura da sentença ou parágrafo. As sentenças 44 e 45 são exemplos de definições ruins.

[74] Os algoritmos de aprendizagem de máquina (SVM) têm como objetivo a determinação de limites de decisão que produzam uma separação ótima entre classes por meio da minimização dos erros (Vapnik, 1995). O SVM consiste em uma técnica computacional de aprendizado para problemas de reconhecimento de padrão. Introduzida por meio da teoria estatística de aprendizagem por Vapnik (1995), essa classificação é baseada no princípio de separação ótima entre classes, tal que se as classes são separáveis, a solução é escolhida de forma a separar o máximo as classes. (NASCIMENTO et al., 2009, p. 2080)

[75] *Ranking SVM is a typical method of learning to rank*. (CAO et al., 2006).

Os autores explicam que já as definições indiferentes estão entre a boa e a ruim. As sentenças de 40 a 43 são modelos dessa categoria (XU et al. 2005).

37	Linux is an open source operating system that was derived from UNIX in 1991.
38	Linux is a UNIX-based operating system that was developed in 1991 by Linus Torvalds, then a student in Finland.
39	Linux is a free Unix-type operating system originally created by Linus Torvalds with the assistance of developers around the world.
40	Linux is a command line based OS.
41	Linux is the best-known product distributed under the GPL.
42	Linux is the platform for the communication applications for the dealer network.
43	Linux is a Unicode platform.
44	Linux is an excellent product.
45	Linux is a threat to Microsoft's core businesses.

CORPUS E METODOLOGIA

A partir de um conjunto de documentos, foram coletados todos os parágrafos que correspondiam às regras de heurísticas propostas (Quadro 7), tendo como saída os candidatos a definições. A descrição dessa etapa é apresentada a seguir.

Todas as sentenças com base NP (*noun phrase*) foram etiquetadas (analisadas gramaticalmente) e, dentro das etiquetas, o termo foi identificado com base em alguma regra[76]. Em seguida, candidatos a definições foram extraídos de acordo com os padrões apresentados no Quadro 7.

Quadro 7 – Padrões definitórios

<term> is a|an|the *
<term>, *, a|an|the *
<term> is one of *

Fonte – XU et al. (2005).

[76] Regra 1. <*term*> is the first Base NP of the first sentence. Regra 2. Two Base NPs separated by 'of' or 'for' are considered as <*term*>. For example, 'Perl for ISAPI' is the term from the sentence "Perl for ISAPI is a plug-in designed to run Perl scripts"

Destacamos que o '*' significa uma *string* que pode conter uma ou mais palavras e '|' significa "ou".

A segunda etapa do processo consiste em classificar os candidatos à definição, com base nos critérios mencionados anteriormente. Para resolver essa questão da classificação, foi adotada a abordagem estatística de aprendizado de máquina. Os candidatos à definição foram etiquetados e usados como dados de treino. Esse processo é descrito formal e detalhadamente em Xu et al. (2005).

RESULTADOS E AVALIAÇÃO

Como a avaliação realizada desse trabalho diz respeito aos classificadores e apresenta outras medidas de avaliação diferentes dos trabalhos descritos anteriormente, não iremos abordá-la no escopo dessa revisão. Apenas ressaltamos que os autores conduziram experimentos, considerando o ranqueamento de definições nos níveis de sentença e de parágrafo, além disso, avaliaram se os modelos de treinamento são independentes de domínio. E os resultados mostraram, de acordo com os autores, que a metodologia é adequada para essas hipóteses lançadas (XU et al. 2005).

COMENTÁRIOS

O trabalho de Xu e equipe, desenvolvido no âmbito da Microsoft, traz para nós uma ideia da necessidade de formalização desse tipo de conhecimento no desenvolvimento de sistemas de busca. A proposta de ranqueamento de um conjunto de definições é um passo que vai além da tarefa de recuperação "desordenada" de contextos definitórios. Outro ponto interessante é o emprego de técnicas de aprendizado de máquina para a classificação automática do conjunto de fragmentos, a partir de um grupo de *features* fornecido ao sistema. Finalizando, a pesquisa apresenta um aspecto distinto de outros trabalhos que vimos até agora: o fato de o grupo privilegiar apenas três padrões definitórios (apresentados no Quadro 7), pois o intuito era testar a metodologia adotada.

6.2.5 ExContext

DESCRIÇÃO GERAL E LÍNGUA DE APLICAÇÃO

Wendt (2010), em sua pesquisa de mestrado defendida na Faculdade de Informática da PUC/RS, propôs, implementou e avaliou um conjunto de heurísticas para a extração automática de potenciais contextos definitórios em textos de língua portuguesa anotados morfossintaticamente. Seu trabalho faz parte de um projeto maior, inserido no grupo de pesquisa em PLN da PUC-RS, que trata da construção automática de ontologias, sendo a extração de contextos definitórios uma das etapas previstas. Como afirma Wendt (2010, p.24), "uma motivação para o desenvolvimento deste trabalho é que a etapa de extração de contextos definitórios complementa o trabalho de construção automática de ontologias (...)"

CONSTRUCTO

Na identificação dos padrões definitórios, o pesquisador utilizou um concordanciador a fim de visualizar o contexto à direita e à esquerda dos termos detectados previamente. Feito isso, foi realizada uma análise manual dos contextos a fim de observar quais contextos eram realmente definitórios. Esses foram agrupados de acordo com a classificação apresentada no Quadro 8.

Quadro 8 – Classificação de contextos definitórios em Wendt (2010, p.40-43).

Padrão	Heurística
Sintático	Ser
Tipográfico	: / ( )
Verbal	formar; compor; constituir; denotar; mostrar; representar; definir; consistir; indicar; significar; simbolizar; caracterizar; conter; apresentar
Indicativo	conhecido como; isto é; reconhecido como

Fonte – WENDT (2010, p.40-43). Adaptado pela autora.

No padrão sintático, são recuperados os contextos que têm o termo seguido diretamente do verbo "ser" e suas flexões. O mesmo ocorre no padrão tipográfico. Já no padrão verbal, são recuperados os contextos que contenham o termo e o verbo (e suas flexões), independente do lugar que ocupam na sentença. No padrão indicativo se procedeu de modo análogo ao padrão verbal.

CORPUS E METODOLOGIA

Como *corpus* de estudo, foi utilizado um conjunto de textos da área de Geologia Geral, com aproximadamente um milhão de palavras, o qual é composto por artigos (119 textos - 815.381 palavras), teses (9 textos - 110.788 palavras) e dissertações (9 textos - 88.528 palavras) do domínio eleito.

A anotação linguística do *corpus* foi feita pelo *parser* PALAVRAS[77], e a sua saída foi armazenada num arquivo *xml*, o qual contém todas as palavras do documento e suas respectivas classificações morfossintáticas.

Em seguida foi realizada a extração de candidatos a termos desse mesmo *corpus*, utilizando para isso a ferramenta E*x*ATO*LP*[78], que funciona aproximadamente da seguinte forma: um *corpus* anotado linguisticamente é submetido à ferramenta que, por sua vez, extrai automaticamente todos os sintagmas nominais (SNs), classificando-os de acordo com o número de *tokens* que o compõem.

Após a extração de candidatos a termos pela ferramenta, o passo foi validar quais candidatos seriam, de fato, termos da Geologia. A fim de cumprir essa tarefa, dois glossários de referência, da MINEROPAR[79] e outro produzido pela UNB[80], e mais a Wikipédia foram utilizados. Além da validação do termo, a tarefa de avaliação consistiu em recuperar e armazenar definições dos termos encontrados no *corpus* e que foram encontrados também em uma das fontes citadas acima. De uma lista inicial de candidatos a termos com 4.889, chegou-se a uma relação de 2.123 termos que possuem definição.

O outro *corpus* composto por oito textos de Química Geral, provenientes de obras didáticas (ATKINS e RUSSEL[81]) também fora anotado pelo PALAVRAS e, ao final de todo o processo, foram encontrados 295 termos que apresentaram definição no próprio *corpus*.

RESULTADOS E AVALIAÇÃO

A avaliação das heurísticas consistiu em uma etapa manual, realizada por especialistas da área da Química e de Geologia. Isso serviu para contrastar os resultados

[77] Disponível em: <visl.sdu.dk/~eckhard/pdf/PLP20-amilo.ps.pdf>. Acesso em ago. 2013.
[78] Disponível em: <www.inf.pucrs.br/~ontolp/exato.php>. Acesso em ago. 2013.
[79] Disponível em: <www.mineropar.pr.gov.br/>. Acesso em ago. 2013.
[80] Disponível em: <www.unb.br/ig/glossario/>. Acesso em jul. 2014. Até o encerramento deste texto, a página do glossário estava inacessível.
[81] Esses textos foram compostos a partir de uma seleção dos capítulos mais relevantes para o conhecimento da área de Química Geral. O *corpus* foi desenvolvido pela equipe TEXTQUIM do Instituto de Letras e pela equipe da Área de Educação Química (AEQ), ambas da Universidade Federal do Rio Grande do Sul (WENDT, 2010 p. 36).

manuais com os automáticos gerados pelas heurísticas. Os contextos extraídos pelo sistema eram classificados pelos especialistas em bom, potencial e ruim.

O *corpus* de teste de Geologia geral foi formado por nove dissertações e nove teses. Com o uso do E*x*ATO*LP*, foram extraídos os 10 termos unigramas, 10 termos bigramas e 10 termos trigramas mais frequentes do *corpus* de teste e que estavam presentes nos glossários de referência. No total, foram encontrados no *corpus* 1.498 ocorrências desses 30 termos.

Com o uso das heurísticas propostas nesse trabalho, a quantidade de contextos caiu de 1.498 para 552. Desse total, na avaliação dos especialistas em Geologia, foram considerados 37 como bons e 48 como potenciais, totalizando 85 contextos relevantes. Segundo o autor, isso representa precisão de 15,4%. Além disso, os especialistas avaliaram que dessas 1.498 ocorrências, 152 eram consideradas contextos definitórios (47 contextos classificados como bons e 107 como potenciais), o que representa uma cobertura de 55,9%, quando considerado o total de contextos definitórios recuperados pelas heurísticas e que coincidiam com os apontados pelos especialistas. O detalhamento da avaliação por heurística é fornecido na Tabela 9.

Tabela 9 – Avaliação por heurística, do domínio de Geologia, em Wendt (2010, p.51).

Heurística	Bom	Potencial	Ruim	Precisão	Cobertura
Formar	18	3	41	3,8%	13,8%
Ser	6	12	62	3,3%	11,8%
Definir	4	6	30	1,8%	6,6%
Apresentar	3	7	90	1,8%	6,6%
Caracterizar	2	4	35	1,2%	4%
Constituir	4	0	35	0,7%	2,6%
Representar	0	4	55	0,7%	2,6%
Compor	0	4	29	0,7%	2,6%
Indicar	0	3	37	0,5%	2%
Consistir	0	2	6	0,4%	1,3%
()	0	2	16	0,4%	1,3%
:	0	1	5	0,1%	0,7%
Mostrar	0	0	15	0%	0%
Denotar	0	0	1	0%	0%
Total	37	48	457	15,4%	55,9%

Fonte – WENDT (2010, p.51). Adaptado pela autora.

Na área de Química Geral, foram selecionados randomicamente 10 termos bigramas e 10 termos trigramas, sendo que para esse grupo foram detectadas 246 ocorrências. Os especialistas na área classificaram tais contextos em bom, potencial e ruim.

Na avaliação humana, 122 dos 246 contextos foram considerados bons ou potenciais, o que revela que praticamente metade dessas ocorrências é composta por contextos definitórios. Empregadas as heurísticas, foram recuperados 102 contextos. Desses, 58 foram avaliados como válidos pelos especialistas (entre contextos classificados como bons e potenciais). Esses dados resultam em uma precisão de 59,6% e cobertura de 47,5%. Na Tabela 10, pode-se observar o detalhamento da avaliação por heurística.

Tabela 10 – Avaliação por heurística do domínio de Química Geral, em Wendt (2010, p.55).

Heurística	Bom	Potencial	Ruim	Precisão	Cobertura
Ser	12	5	7	16,6%	13,9%
Formar	4	6	7	9,8%	8,2%
Chamar	6	1	1	6,8%	5,7%
Consistir	5	0	2	4,9%	4,1%
Mostrar	1	2	4	2,9%	2,6%
Representar	1	2	5	2,9%	2,6%
Constituir	2	0	0	2%	1,6%
Caracterizar	2	0	3	2%	1,6%
()	1	1	8	2%	1,6%
Significar	1	1	0	2%	1,6%
Conhecido como	1	0	0	1%	0,8%
Definir	1	0	1	1%	0,8%
Isto é	1	0	0	1%	0,8%
:	0	1	1	1%	0,8%
Apresentar	0	1	2	1%	0,8%
Indicar	0	0	3	0%	0%
Total	38	20	44	56,9%	47,5%

Fonte – WENDT (2010, p.55). Adaptado pela autora.

COMENTÁRIOS

O trabalho de Wendt (2010) pode ser considerado como uma iniciativa atual da área de PLN na recuperação de contextos definitórios para fins de construção de ontologia

no cenário do Português do Brasil. O trabalho foi ambicioso no sentido de contemplar três tipos de contextos definitórios distintos – sintático, verbal e tipográfico e, além do algoritmo feito para a tarefa, foi ainda apresentado um protótipo em Java, o *ExContext*, para a extração de contextos definitórios por qualquer pessoa que se interessar pela tarefa. Como afirma o pesquisador: "pode-se concluir que os métodos empregados neste trabalho fornecem uma grande redução de material a ser analisado" (WENDT, 2010, p.57). Embora os melhores resultados exibidos na avaliação (considerados os contextos bons e potenciais) girem em torno de 50%, por meio de heurísticas mais refinadas a partir de um estudo mais apurado desses padrões, aumenta-se a possibilidade de melhorar a precisão e a cobertura na avaliação desse mesmo conjunto.

6.2.6 Ecode

DESCRIÇÃO GERAL E LÍNGUA DE APLICAÇÃO

Em sua pesquisa de Doutorado em Ciências da Linguagem e Linguística Aplicada, realizado na *Universitat Pompeu Fabra* (UPF), a investigação de Alarcón (2009) teve por objetivo propor uma metodologia para a extração automática de contextos definitórios em um *corpus* terminológico de língua espanhola com o intuito de auxiliar no processo terminográfico. Seu trabalho, como afirma o autor, constitui uma continuação de uma investigação realizada no *Grupo de Ingeniería Lingüística* (GIL),[82] da Universidade Nacional Autônoma do México, em nível de licenciatura, onde o tema central foi a análise linguística de contextos definitórios em textos especializados em espanhol. A metodologia de extração de contextos definitórios baseou-se em padrões verbais (ALARCÓN, 2003; SIERRA e ALARCÓN, 2003). Ao final, o sistema construído propõe, além da extração de ocorrências de padrões definitórios de base verbal, filtros automáticos de modo a descartar aqueles contextos que não são definitórios de fato e a identificar os elementos constitutivos dos contextos definitórios.

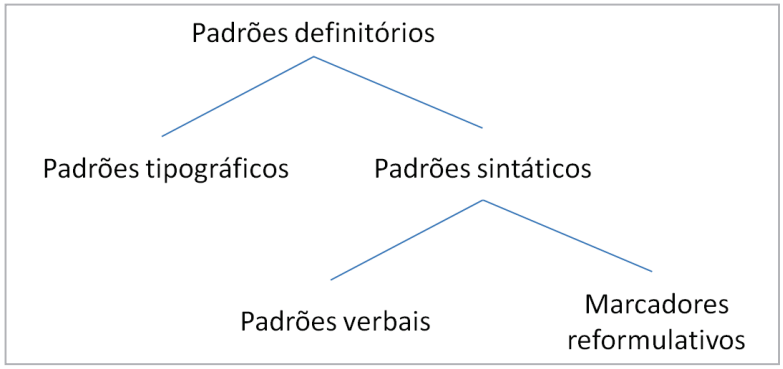
CONSTRUCTO

A partir da descrição feita por Alarcón (2003) e Sierra e Alarcón (2003), é proposta uma tipologia de padrões definitórios que indica explicitamente certo tipo de informação conceitual e que, além disso, constitui uma chave essencial no processo de

[82] <www.iling.unam.mx/>. Acesso em fev. 2014.

reconhecimento de contextos definitórios de modo automático (ALARCÓN, 2009, p.125). No Esquema 7, a tipologia de padrões definitórios é apresentada.

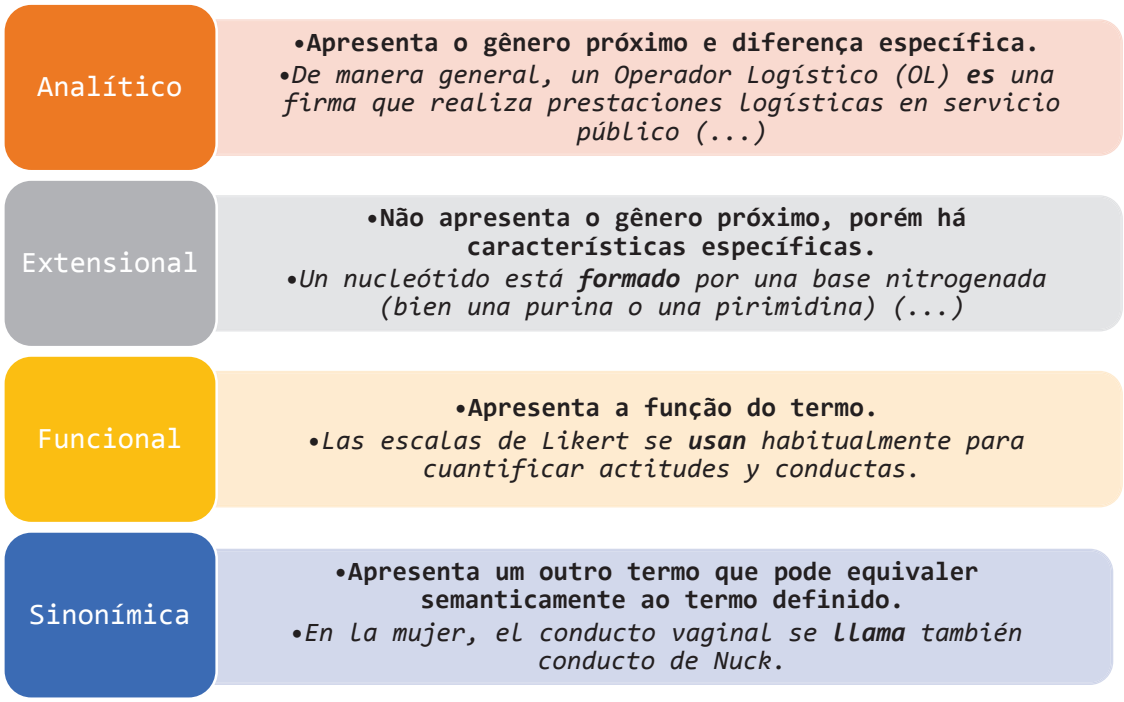
Esquema 9 – Tipologia de padrões definitórios, de acordo com Alarcón (2009, p.127).



Fonte – ALARCÓN (2009, p.127). Adaptado pela autora.

A fim de delimitar o escopo da pesquisa, foram eleitos os padrões definitórios sintáticos, de núcleo verbal, classificados e exemplificados na pesquisa segundo os dados da Figura 4, extraídos de Alarcón (2009, p.135-138).

Figura 4 – Padrões definitórios sintáticos de núcleo verbal, em ALARCÓN (2009, p.135-138).



Fonte – ALARCÓN (2009, p.135-138). Adaptado pela autora.

A gramática definitiva de padrões definitórios proposta em Alarcón (2009), que é exibida no Quadro 9, contempla os seguintes verbos e apresenta seis campos. São eles:

a) **Contexto definitório (CD):** o primeiro campo apresenta a classificação do padrão definitório em analítico (A), extensional (E), funcional (F) e sinonímico (S);

b) **Lema:** o segundo campo apresenta a forma canônica do verbo núcleo do padrão definitório;

c) **Raiz[83]:** o terceiro campo apresenta o radical do verbo que será considerada na busca;

d) **Distância (Dist):** o quarto campo apresenta a distância máxima possível entre a raiz e o nexo. O "0" (zero) representa que o nexo deverá constar imediatamente após a raiz, enquanto que o "-" significa que não há distância máxima entre eles.

e) **Nexo:** o quinto campo apresenta o nexo, o qual funciona como um elemento de ligação entre o termo e a definição.

f) **Padrão Contextual (PC):** o sexto campo apresenta a posição do termo em relação ao verbo definitório, que pode ser: *izquerda* (I) – quando o termo aparece à esquerda do padrão verbal; *derecha* (D) – quando o termo aparece à direita do padrão verbal; *nexo* (N) – quando o termo aparece entre o padrão verbal e o nexo.

[83] O termo "raiz" do espanhol equivale a "radical" no português.

Quadro 9 – Gramática de padrões definitórios, em ALARCÓN (2009, p.159).

CD	Lema	Raiz	Dist	Nexo	PC
A	**Ser**	(es/son)	0	determinante	I
	Caracterizar	caracteriz	-	como, por	I
	Concebir	conc(e|i)b	-	como	I N
	Considerar	consider	-	como	I N
	Describir	describ	-	como	I N
	Definir	defin	-	como	I N D
	Entender	ent(ie|e)nd	-	como	I N D
	Conocer	conoc	-	como	I N D
	Denominar	denomin	-	ø, como	D
	Llamar	llam	-	ø, como	D
	Nombrar	nombr	-	ø, como	D
E	**Comprender**	comprend	0	ø	I
	Contener	cont(ien|en|uv)	0	ø	I
	Incluir	inclu(í|i|y)	0	ø	I
	Integrar	integr	0	ø	I
	Constar	const(a|e|ó)	-	de	I
	Contar	c(ue|o)nt(a|e|á|é|ó)	-	con	I
	Formar	form	-	de, por	I
	Componer	comp(on|us|uest)	-	de, por	I
	Constituir	constit	0	de, por	I
F	**Permitir**	permit	0	ø	I
	Encargar	encarg	-	de	I
	Consistir	consist	-	en	I
	Funcionar	funcion	-	como, para	I
	Ocupar	ocup	-	como, para	I
	Servir	s(i|e)rv	-	como, para	I
	Usar	us	-	como, en, para	I N
	Emplear	emple	-	como, en, para	I N
	Utilizar	util	-	como, en, para	I N
S	**Conocer**	conoc	0		D
	Denominar	denomin	0		D
	Llamar	llam	0		D
	Nombrar	nombr	0		D

Fonte – ALARCÓN (2009, p.159).

A metodologia utilizada para a construção do sistema partiu do *Corpus* Técnico do IULA, em espanhol, o qual foi etiquetado morfossintaticamente com POS – *Part of Speech*. Nesse *corpus*, estão inclusos documentos das áreas de Direito, Genoma, Economia, Meio Ambiente, Medicina, Informática e Língua Geral (BACH et al, 1997) e apresenta 1.091.946 *tokens,* sendo 38.247 orações.

Todo o processo de extração de contextos definitórios foi implementado mediante uma série de *scripts* que, em seu conjunto, formam o ECODE (Extrator de Contextos Definitórios). Em linhas gerais, a construção do sistema está alicerçada em três importantes etapas:

a) Aquisição de ocorrências de padrões definitórios

A partir do *corpus* de estudo, foram feitas buscas automáticas por instâncias de padrões verbais definitórios. Uma vez encontrados os padrões, foi feita anotação *xml* do padrão **<pvd>,** da sequência que veio anteriormente ao padrão **<izq>**, e da sequência que veio após ao padrão **<der>**. Finalmente, nos casos em que o padrão verbal incluía um nexo como a partícula "como" (ex.: "X se define <u>como</u> Y") era etiquetado com **<nx>**. Podemos ver em 46 um exemplo da anotação.

46	**<izq>**El metabolismo**</izq><pvd>**puede definir se**</pvd><nx>**en términos generales como**</nx><der>**la suma de todos los procesos químicos (y físicos) implicados.**</der>**

b) Filtro de contextos não relevantes

Após essa etapa, os candidatos a contextos definitórios anotados foram submetidos a um filtro responsável por eliminar falsos contextos definitórios, com base em uma gramática de restrições, constituída por sequências sintáticas ou partículas gramaticais que poderiam se manifestar em determinadas posições na sentença – antes do verbo, após o verbo, ou ainda entre o verbo e o nexo. A relação na íntegra das expressões utilizadas para filtrar os candidatos a contextos definitórios é apresentada no Quadro 10.

Quadro 10 – Filtro de contextos não relevantes, em ALARCÓN (2009, p.167).

Posição		Regra
Esquerda	1	para <pvd>
Nexo	2	</vd>.* verbo conjugado .*</nx>
	3	</vd>.*? se .*?<nx>
	4	</vd>.*? tanto .*?<nx>
	5	</vd>.*? sino .*?<nx>
	6	</vd> , <nx>
	7	así <nx>
	8	cerca <nx>
	9	parte <nx>
	10	partir <nx>
	11	más <nx>
	12	menos <nx>
	13	mientras (que,|que)<nx>
	14	no <nx>
	15	poco <nx>
	16	poco más <nx>
	17	(que,|que)<nx>
	18	sino <nx>
	19	tales <nx>
	20	tal <nx>
	21	y <nx>
	22	ya <nx>
Direita	23	<\pvd> antes
	24	<\pvd> cuan
	25	<\pvd> para
	26	<\pvd> si
	27	<\pvd> se
	28	<\pvd> verbo Conjugado

Fonte – ALARCÓN (2009, p.167).

Dessa forma, a partir das regras elaboradas, foi possível excluir sentenças como a apresentada no exemplo 47 (ALARCÓN, 2009, p.168).

47	Para <**pvd**> <**vd**>entender</**vd**> <**nx**>como</**nx**> </**pvd**> funciona el ADN, es necesario conocer algo sobre su estructura y organización.

c) Identificação automática de elementos constitutivos do contexto definitório

Com o objetivo de identificar automaticamente as possíveis posições da definição e do termo nos candidatos a contextos definitório, foi utilizada uma árvore de

decisão baseada em determinadas heurísticas, por meio de inferências lógicas e dependendo de cada padrão verbal. Por meio desse processo, é possível detectar, por exemplo, em contextos definitórios com o verbo "significar", que o termo pode se manifestar apenas à esquerda do verbo, e a definição constará à direita do verbo (*T significa D*).

RESULTADOS E AVALIAÇÃO

Como *corpus* de avaliação, foi usado o *sub-corpus* do IULA da área de Medicina. Como o objetivo da avaliação foi validar o sistema que tinha como núcleo verbos definitórios com suas respectivas regras, o primeiro passo foi buscar os lemas (formas canônicas) desses verbos definitórios no *corpus* de estudo. A quantidade de verbos encontrados variou de um a no máximo 250, sendo que os nexos eram inclusos (quando o verbo exigia), considerando uma janela de 15 palavras entre o verbo e ele.

Em seguida, foi feita uma análise manual a fim de verificar quais eram contextos definitórios de fato. Isso foi feito por avaliadores envolvidos com Terminologia, cada qual com uma parte do *corpus*. O único requisito solicitado era que considerassem como contexto definitório aqueles casos em que há explicitamente um termo e uma definição. Como resultado dessa análise manual, foi obtido um *corpus* de avaliação, o qual apresenta etiquetas para fragmentos que são CDs e para fragmentos que não são CDs. Esse *corpus* de avaliação serviu de *input* para o ECODE, sobre o qual foram aplicados os índices de cobertura e precisão. Os resultados globais sem a gramática de restrições estão exibidos na Tabela 11.

Tabela 11 - Avaliação global em Alarcón (2009, p.203).

Precisão	Cobertura
0.53	0.79
1783 de 3309	1783 de 2254

Fonte – ALARCÓN (2009, p.203).

Os resultados da avaliação global sugerem que o algoritmo desenvolvido obteve um índice médio quanto à precisão (0.53) e um índice melhor quanto à cobertura (0.79).

Em uma avaliação pormenorizada, podemos observar os resultados referentes a cada um dos padrões, como mostram as Tabelas 12, 13, 14 e 15.

Tabela 12 – Avaliação dos padrões analíticos em Alarcón (2009, p.211).

Padrões Analíticos		
Padrão verbal	**Precisão**	**Cobertura**
(es|son)+det	0,6	0,78
caracterizar como	0,37	0,75
caracterizar por	0,66	1
concebir como	0,5	0,66
considerar como	0,57	0,8
describir como	0,65	0,85
definir como	0,85	0,87
entender como	0,69	0,69
conocer como	0,48	0,95
denominar como	0,4	0,66
llamar como	1	1
denominar	0,62	0,89
llamar	0,62	0,78
nombrar	0,14	1

Fonte – ALARCÓN (2009, p.211).

Tabela 13 – Avaliação dos padrões funcionais em Alarcón (2009, p.213).

Padrões Funcionais		
Padrão verbal	**Precisão**	**Cobertura**
permitir	0,68	0,73
encargar de	0,29	0,74
consistir en	0,44	0,79
funcionar como	0,63	0,9
funcionar para	0,5	0,66
ocupar como	0,2	1
ocupar para	0,5	0,66
servir como	0,66	0,8
servir en	0,25	1
servir para	0,47	0,77
usar como	0,84	1
usar en	0,37	0,68
usar para	0,73	0,84
emplear como	0,42	1
emplear en	0,13	0,85
emplear para	0,29	0,82
utilizar como	0,38	0,88
utilizar en	0,33	0,9
utilizar para	0,46	0,76

Fonte – ALARCÓN (2009, p.213).

Tabela 14 – Avaliação dos padrões extensionais em Alarcón (2009, p.212).

Padrões Extensionais		
Padrão verbal	**Precisão**	**Cobertura**
comprender	0,25	0,82
contener	0,64	0,73
incluir	0,42	0,54
integrar	0,18	0,69
constar de	0,73	0,91
contar con	0,26	0,61
formar de	0,37	0,83
formar por	0,83	0,76
componer de	0,6	0,75
componer por	0,72	0,85
constituir de	0,33	1
constituir por	0,4	0,81

Fonte – ALARCÓN (2009, p.212).

Tabela 15 – Avaliação dos padrões sinonímicos em Alarcón (2009, p.214).

Padrões Sinonímicos		
Padrão verbal	**Precisão**	**Cobertura**
Conocer también	0,77	0,8
Denominar también	0,72	0,94
llamar también	0,79	0,81

Fonte – ALARCÓN (2009, p.214).

Houve padrões que apresentaram excelente precisão e cobertura, como "llamar como", "usar como", "constar de", outros padrões apresentaram cobertura absoluta, porém baixa precisão, como "ocupar como", "nombrar" e "constituir de". E, por fim, não houve nenhum caso em que a precisão fosse superior à cobertura.

Em seguida, foi feita outra avaliação, considerando as seguintes restrições:

a) Raízes dos verbos definitórios: 3ª pessoa do singular ou plural, no presente, infinitivo ou particípio;
b) Distância entre o verbo e o nexo em uma janela de 15 palavras;
c) Padrões Contextuais (PC).

Após a aplicação das restrições, obteve-se a taxa de 0.54 para precisão e 0.74 para a cobertura na avaliação global. Ainda com alguns ajustes, o resultado pôde ser melhorado em 0.55 para precisão e 0.77 na cobertura.

COMENTÁRIOS

A investigação realizada por Alarcón (2009) reúne um percurso comum às demais pesquisas: a partir da descrição dos contextos definitórios feita previamente, os dados obtidos foram implementados e, a partir do sistema implementado, seu desempenho foi julgado.

Contudo, sua relevância para a presente pesquisa reside no fato de que sua investigação está alinhada com a nossa proposta, ou seja, parte da necessidade de construção de recursos úteis para a atividade de construção de produtos terminológicos. Mais do que isso, seu trabalho apresenta uma classificação e descrição linguística fundamentada e bem refinada, o que é de grande valia, uma vez que a descrição é o cerne do nosso estudo.

O sistema ECODE possibilita: a) detectar os padrões verbais, b) filtrar quais deles apresentam contextos definitórios não relevantes com base em uma gramática de

restrições e c) identificar os elementos que constituem os contextos definitórios. Todos esses passos foram expostos com detalhes de modo que este trabalho pode ser replicado e complementado em investigações de outros grupos de pesquisa.

6.2.7 Extração de contextos definitórios em outras pesquisas

Além desses trabalhos resenhados, há outros que tiveram como objetivo também a descrição e extração de contextos definitórios. Saggion (2004) e Saggion e Gaizauskas (2004) focam na extração de definições para sistemas automáticos de pergunta e resposta em domínios especializado e geral. Os autores partem do desafio de encontrar termos e definições em uma grande coleção de textos a partir de um método que utiliza fontes externas já disponíveis como a *WordNet*[84] (MILLER, 1995), a Enciclopédia Britânica[85] e o buscador Google. Ao encontrar o termo e sua respectiva definição nesses recursos *on-line*, são armazenados os chamados termos secundários, os quais ocorrem no *definiendum* do termo de busca com a finalidade de serem empregados na recuperação dos contextos definitórios em grandes quantidades de textos.

Outro trabalho que podemos citar é a pesquisa desenvolvida por Malaisé et al. (2005), que tratou da extração de contextos definitórios em francês com vistas à elaboração de ontologias diferenciais, as quais se entendem como estruturas terminológicas hierárquicas normalizadas que constituem o primeiro passo na construção de qualquer ontologia. (MALAISÉ et al., 2005). A metodologia se baseou em *corpus* com textos de gêneros diversos. A extração de contextos definitórios seguiu por meio dos seguintes padrões: metalinguístico (denominar, definir); marcadores metalinguísticos (definição, nome) em combinação com verbos (referir, utilizar); marcadores discursivos (isto é, em outros termos) e signos de pontuação (especificamente parênteses). A partir das buscas por esses padrões no *corpus*, foram detectados o termo definido (principal) e o termo secundário, o qual poderia ser classificado de acordo com sua relação com o termo principal (sinônimo, hiperônimo, merônimo e etc.).

Já Navigli e Velardi (2007) apresentam o sistema *web GlossExtractor*. Sua principal função é extrair candidatos à definição, a partir de glossários, documentos ou páginas específicas oriundos da *web*. Sinteticamente, o *GlossExtractor* parte de uma lista

[84] <wordnet.princeton.edu/wordnet/>. Acesso em fev. 2014.
[85] <www.britannica.com/>. Acesso em fev. 2014.

de termos fornecida pelo usuário, a qual serve de busca primeiramente em glossários *on-line* indexados previamente ao sistema. Em seguida, para cada um dos termos da lista, são gerados alguns padrões de busca como *<term> is a, <term> defines*, entre outros. Cada sequência conforma uma expressão de busca lançada no Google, sendo que, para cada uma, são consideradas as cinco primeiras páginas, as quais são convertidas no formato *txt*. As duas etapas seguintes consistem no tratamento de ruído encontrado no resultado dessas buscas. A primeira se refere a um filtro estilístico que busca as definições mais canônicas, ou seja, as formadas por gênero próximo e diferença específica. Foi possível construir esse filtro graças a um algoritmo de aprendizagem automática que permitiu gerar uma árvore de decisão para efetuar esse processo. A outra etapa diz respeito a um filtro que elimina definições não pertinentes ao domínio de interesse, a partir de um modelo probabilístico fornecido com a análise de domínio da terminologia de entrada.

O tema também é de interesse de pesquisa para cientistas da informação, como podemos observar no trabalho de mestrado de Oliveira Jr. (2012), que propôs uma gramática de padrões definitórios com base em Alarcón (2009) e Kamikawachi (2009), a partir de estudo com *corpus* formado por teses e dissertações da própria área da Ciência da Informação. Tais padrões foram classificados em quatro estruturas diferentes: a) EATED (Expressão linguística 01, Autor, Termo, Expressão Linguística 02, Definição)[86]; b) AETED (Autor, Expressão linguística 01, Termo, Expressão Linguística 02, Definição)[87]; c) TED (Termo, Expressão linguística 01, Definição)[88]; d) ETED (Expressão linguística 01, Termo, Expressão linguística 02, Definição)[89].

De acordo com o levantamento realizado na investigação, as estruturas, da mais frequente para a menos frequente na amostra do *corpus*, são: TED (62%), AETED (14%), EATED (12%) e ETED (04%). Os resultados gerados na pesquisa serão úteis na área de Descoberta de Conhecimento de Textos e Extração de Informação (OLIVEIRA JR., 2012, p.100).

---

86 Ex.: "Para Flory (2005), empregabilidade é a qualidade de manter-se no mercado, ser desejado pelos alvos e coerentes com a missão." (OLIVEIRA JR., 2012, p.86).

87 Ex.: "Wilson (2006) afirma que a gestão do conhecimento é uma extensão dos conceitos da gestão da informação." (OLIVEIRA JR., 2012, p.88).

88 Ex.: "O conhecimento é também o resultado dos relacionamentos que a organização manteve ao longo do tempo com seus clientes, fornecedores e parceiros." (OLIVEIRA JR., 2012, p.89).

89 Ex.: "O objetivo da recuperação, dada uma pergunta formalizada por descritores que a definem corretamente, é que o sistema de informação providencie a comparação desses com aqueles que descrevem o documento e obtenha as referências bibliográficas que atendem à pergunta em questão." (OLIVEIRA JR., 2012, p.89).

Por fim, fazemos menção ao método *Definition Extraction*, patenteado em 2008 por Powell, Humphreys e Azzam, da *Microsoft Corporation*[90], que tem como propósito a extração de definições em documentos, com base na identificação de *cue-phrase* (ex.: "is-a"), seguido de técnicas de processamento linguístico mais profundo.

Nesse percurso da revisão bibliográfica de contextos definitórios, observamos que houve um interesse nesse tópico nos últimos 10 anos, especialmente pelas áreas que de alguma forma estão envolvidas com Recuperação de Informação e Terminologia. Hoje, com a crescente quantidade de informações ao nosso redor, faz-se necessário que se localize rapidamente a informação desejada, e uma das informações requeridas pode ser, entre outras, uma definição. Assim, o tópico de contexto definitório dialoga com áreas que têm como objeto: ensino à distância, construção de dicionários e glossários, elaboração de ontologia, sistemas de perguntas e respostas e sistemas de busca, sempre com a intenção de se recuperar brevemente uma dada informação por meio de sentença(s), parágrafos e/ou documentos.

[90] Patente Nº: US 7376551

PARTE 3

METODOLOGIA E ANÁLISE DA PESQUISA

7. METODOLOGIA NA IDENTIFICAÇÃO DE CONTEXTOS DEFINITÓRIOS

Aquilo que se vê depende do lugar em que foi visto, e das outras coisas que foram vistas ao mesmo tempo.
Clifford Geertz

Nesta seção, apresentamos o percurso seguido na identificação e extração de contextos definitórios, desde a caracterização do *corpus* de estudo, o qual será nossa amostra de língua/linguagem, até a técnica empregada para o reconhecimento dos candidatos a contextos definitórios. Ressaltamos que a investigação centra-se na descrição de padrões definitórios, ou seja, na análise de determinadas expressões linguísticas que indicam que um fragmento ou sentença contém um termo e sua respectiva definição ou explicação.

Assim, primeiramente o *corpus* de trabalho é apresentado por meio de sua descrição qualitativa e quantitativa (Estágio 1). É salientado também como foi feita a escolha dos textos que comporiam o *corpus*, uma vez que era demasiadamente grande (mais de 75 mil textos, entre artigos, teses, dissertações, resenhas, resumos, editoriais e etc.). Em seguida, os padrões léxico-sintáticos utilizados na investigação são pormenorizados (Estágios 2 e 3). Por fim, o método de investigação quanto à identificação dos contextos definitórios é descrito (Estágios 4 e 5).

Como observado na revisão bibliográfica, algumas propostas foram realizadas quanto à classificação de contextos definitórios (DOCUMENTED, 2005 e ALARCÓN, 2009). Nesta investigação, como ponto de partida, tomaremos como análise os padrões léxico-sintáticos que têm o verbo como núcleo, pois são padrões frequentes e vêm sendo investigados até o momento, mas que, ao mesmo tempo, ainda carecem de uma descrição mais detalhada no caso do português.

Dessa forma, partimos do rol de padrões verbais do espanhol proposto por Alarcón (2009) (Figura 4), pois julgamos, pela proximidade entre as línguas, que parte considerável dos verbos elencados em espanhol é frequente em português na formação de contextos definitórios.

7.1 ESTÁGIO 1 - CONFORMAÇÃO DO *CORPUS* DE ESTUDO

O *corpus* de estudo utilizado nesta pesquisa para a análise dos contextos definitórios *in loco* é composto, na sua totalidade, de artigos científicos, tendo em vista que é nesse tipo de texto que, prototipicamente, as ciências e as técnicas fazem circular seus saberes, constituindo-se, portanto, em textos terminológicos.

Ressaltamos, ainda, que todos os textos são em língua portuguesa, variante brasileira, e que, de um modo geral, pertencem a alguma das quatro áreas do conhecimento: Humanas, Exatas, Agrárias e Biológicas. Isso porque não queríamos focar numa única área do conhecimento para não correr o risco de criar padrões linguísticos dependente de área.

Esse *corpus* de estudo empregado faz parte de um *corpus* maior – o Banco do Português, compilado e organizado pelo LAEL/PUC, sob coordenação de Berber Sardinha. O Banco do Português[91] possui mais de 500 mil arquivos, distribuídos em tipos textuais diversos,[92] em português brasileiro, já convertidos, limpos e em formato "txt". A disponibilização desse *corpus* para a pesquisa foi de grande relevância, uma vez que o mesmo já estava pronto, dispensando, portanto, que compilássemos e organizássemos os textos que serviriam de amostra, e melhor do que isso, ter ciência que a constituição do *corpus* foi fruto de um trabalho sério e competente oriundo de um dos maiores centros de pesquisa em Linguística de *Corpus* no Brasil.

Após a obtenção do Banco do Português, o passo seguinte foi realizar a recuperação de arquivos referentes a artigos científicos. Para que isso ocorresse, antes foi necessário estudar a configuração do *corpus* e, em seguida, foi criado e utilizado o algoritmo[93] representado no Quadro 11, o qual atenderia na tarefa de recuperação de artigos científicos nesta pesquisa.

---

[91] Ressalte-se que, ao final da escrita da tese, o Banco do Português, que integra o *Corpus* Brasileiro, já se encontrava disponível em: http://www.sketchengine.co.uk/.

[92] O *corpus* do Banco do Português organiza-se, em pastas, da seguinte forma: acadêmico, culinária, falado, informática, jornalístico, jurídico, legal, literatura, médico, negócios e religioso.

[93] A fim de reconhecer a estruturação prototípica dos artigos científicos do *corpus*, foi possível contar com a colaboração das linguistas Gabriela Aniceto e Mayara Victor Gomes.

Quadro 11 – Algoritmo para a recuperação de artigos científicos.

```
Se FILE ≥ 15 kb
   Se FILE  =  RESUMO e
     INTRODUÇÃO e
     METODOLOGIA ou MATERIAIS E MÉTODOS ou MATERIAL E MÉTODOS e
     RESULTADOS E DISCUSSÃO ou RESULTADOS ou RESULTADOS E DISCUSSÕES ou RESULTADO E DISCUSSÃO e
     CONCLUSÕES ou CONCLUSÕES FINAIS ou CONSIDERAÇÕES FINAIS e
     REFERÊNCIAS ou LITERATURA CITADA
ENTÃO ARTIGO CIENTÍFICO
```

Fonte - Elaborado pela autora.

Primeiramente, o algoritmo se propõe a recuperar arquivos que são iguais ou maiores que 15 kb (para evitar que fossem recuperados resumos que apresentam explicitamente suas seções) e estes devem ter em seu corpo: início de linha (\n), as expressões elencadas, em caixa alta, de "RESUMO" a "REFERÊNCIAS" ou "LITERATURA CITADA," seguidas de quebra de linha (\r). Isso para que de fato só fossem identificadas essas expressões como subtítulos dos artigos, e não no meio do texto, como em:

48	"Assim, a metodologia utilizada neste trabalho é baseada na redução do óxido de estanho pelo carbono, formando vapor do subóxido [equação (3.1)], e depois na oxidação, ou não, do subóxido que condensa formando as nanoestruturas."
49	"Estes resultados indicam que a amostra branca consiste em uma grande quantidade de nanofitas monocristalinas de SnO2, com comprimento da ordem de vários microns."

Dessa forma, também foi possível eliminar do *corpus*, possíveis textos que se caracterizavam como editoriais, resumos, *abstracts*, resenhas e textos muito incompletos. Embora tenha havido êxito em recuperar uma quantidade razoável de textos que seguem essa organização, obviamente que nem todos os artigos apresentam essa sequência, de modo que alguns artigos ficaram de fora do *corpus* de estudo.

Uma particularidade desse *corpus* em relação a outros *corpora* utilizados em trabalhos já desenvolvidos acerca do tema, tais como Alarcón (2009), Wendt (2010), entre

outros, diz respeito à utilização de um *corpus* sem anotação morfossintática. Essa decisão foi tomada devido ao fato de que existiam duas opções que realizam a busca por categorias morfossintáticas: o programa *Unitex* (PAUMIER, 2002)[94], que dispõe de um léxico acoplado ao sistema, responsável por atribuir classificação morfológica às unidades linguísticas do *corpus*; e o recurso *online Sketch Engine* (KILGARRIFF et al., 2004),[95] que possibilita a busca em *corpus* próprio ou em *corpora* disponibilizados na plataforma, de formas diversas (lema, ocorrência, contexto à esquerda e à direita, entre outros).

Tais recursos supriram as necessidades da pesquisa, uma vez que concentramo-nos sobretudo nas tarefas de identificação e descrição dos padrões definitórios, contudo, entendemos que, para a implementação completa dos resultados da pesquisa, seja avaliada a possibilidade de o sistema de extração de contextos definitórios possuir um *tagger*,[96] de modo a otimizar o seu desempenho.

Destaca-se ainda que, em razão de direitos autorais, parte considerável dos textos está incompleta. Apesar de essa característica ser uma limitação do *corpus*, julgamos que não representa um problema, haja vista que a quantidade de textos que será analisada é bem ampla, possibilitando, portanto, o descarte de textos ou sentenças incompletos.

Em termos quantitativos, o *corpus* de estudo apresenta a seguinte configuração (quadro 12).

Quadro 12 – Configuração do *corpus* de estudo .

Tipo Textual	Arquivos	*Tokens*
Artigos	5.732	23.478.507

Fonte - Elaborado pela autora.

É justamente esse conjunto de textos que é considerado nossa amostra de língua/linguagem. Ressalte-se que, embora a quantidade de arquivos e *tokens* elencada no quadro seja relativamente alta, pela sondagem prévia realizada com os textos que temos à disposição e pelos estudos anteriores sobre o tema, observamos que é necessário ter uma quantidade considerável de textos, com o propósito de observar e descrever padrões de contextos definitórios de forma mais minuciosa.

---

94 *Software* desenvolvido na *Université Paris Est-Marne-La-Vallée* (França) por Sébastien Paumier (PAUMIER, 2002). O *Unitex* consiste num conjunto de programas que permite o processamento de grandes quantidades de textos, em diversas línguas (ALMEIDA e VALE, 2008).

95 <http://www.sketchengine.co.uk/>. Acesso em jul. 2014.

96 *Tagger* é um etiquetador que atribui uma categoria morfossintática a cada palavra do *corpus*.

7.2 ESTÁGIO 2 – LEVANTAMENTO DOS PADRÕES DEFINITÓRIOS DO ESPANHOL

Como mencionado no início da seção 7, valemo-nos dos padrões léxico-sintáticos de núcleo verbal, provenientes do trabalho de Alarcón (2009). Isso foi feito devido ao fato de que, por meio da revisão da literatura, foi constatado que não era necessária uma nova classificação de contextos definitórios, uma vez que isso já fora feito satisfatoriamente em boa parte dos trabalhos da área. Isso posto, o passo seguinte foi definir quais grupos verbais seriam analisados.

Considerando os verbos analisados por Alarcón (2009) (expostos no Quadro 9), em seu trabalho são descritos seis verbos classificados como Analítico, a saber: *concebir, entender, conocer, denominar, llamar* e *nombrar*. Essa eleição considerou o fato de que é maior a produtividade dos verbos do grupo Analítico na formação de CDs em comparação aos outros grupos.

7.3 ESTÁGIO 3 – PROCESSO DE EQUIVALÊNCIA DOS PADRÕES DEFINITÓRIOS DO PORTUGUÊS

O passo seguinte foi o processo de equivalência dos lemas dos verbos utilizados no espanhol para o português. Para essa tarefa, nos valemos do recurso *on-line Linguee*[97], buscador de tradução que permite observar a expressão de busca em textos autênticos (a ferramenta pode ser chamada de concordanciador paralelo), nas línguas de partida e chegada.

Por meio desse recurso, examinamos as possíveis equivalências dos lemas em espanhol, para o português, de acordo com o seu contexto de uso. Ao final do processo, propomos as seguintes equivalências para os verbos do espanhol (quadro 13).

Quadro 13 – Proposta de equivalência dos verbos do espanhol.

Analítico
chamar

[97] <www.linguee.es/espanol-portugues>. Acesso em fev. 2014.

conceber
conhecer
denominar
entender
nomear

Fonte - Elaborado pela autora.

A partir desses verbos em português, e com o auxílio do conjugador verbal proveniente do *Novo Dicionário Eletrônico Aurélio* (2009), os radicais dos verbos foram recuperados de modo a servir de *input* na gramática inicial. Eles são descritos no Quadro 14.

Quadro 14 – Radicais dos verbos.

Analítico
cham
conceb
conhec
denomin
entend
nome

Fonte - Elaborado pela autora.

De fato, os seis verbos elencados são recorrentes na língua, uns mais que outros, e podem se apresentar como definitórios ou ainda manifestar outros sentidos. Mais à frente neste trabalho, será possível mensurar a produtividade "definitória" de cada verbo.

7.4 ESTÁGIO 4 – IDENTIFICAÇÃO PRELIMINAR DOS CANDIDATOS A VERBOS DEFINITÓRIOS

Isso feito, realizamos uma busca pelos radicais verbais no *corpus* de estudo, a fim de extrair e armazenar as sentenças que apresentam esses radicais, independentemente de ser um contexto definitório ou não. Todavia, devido à grande quantidade de arquivos e de *tokens* do *corpus* e também pelo fato de o *corpus* apresentar quebras de linha dentro da sentença (ao invés de somente no final dos parágrafos), optou-se por fragmentar a tarefa em duas partes.

Em um primeiro momento, com o intuito de diminuir o tamanho do *corpus* e, por conseguinte, o tempo de processamento, elaboramos a seguinte expressão regular[98] (ER) em um processador de arquivos que, com o suporte de ER, rapidamente busca fragmentos textuais em um ou mais diretórios.

```
.*\n.*\n.*\n.* \b(radical1|radical2|radicalz3|radical_n).*\n.*\n.*\n.*
```

Essa ER busca no *corpus* as ocorrências das expressões candidatas a radicais (\bradical$_n$) com três linhas antes (.*\n.*\n.*\n.*) e depois (.*\n.*\n.*\n.*) do radical. A ER inicial permitiu recuperar fragmentos textuais que apresentam a sequência em diferentes lugares da sentença, independente de ser composta por letras maiúsculas ou minúsculas. A título de ilustração, os fragmentos recuperados a partir da expressão "defin" são exibidos:

50	Defini-se estridor como a presença de respiração ruidosa, resultante da turbulência na passagem do ar por estreitamento da via aérea, que, dependendo da localização, pode ser inspiratório (faringe ou supraglote), bifásico (glote ou subglote) ou expiratório (traquéia ou vias aéreas inferiores)1.
51	antes que a interferência se instale de maneira definitiva e reduza significativamente a produtividade da lavoura.
52	porque é no período vegetativo que, em geral, se estabelecem as relações definitivas da competição entre plantas daninhas e cultivadas.
53	ainda não está bem esclarecida como nos estudos do trato genital, na qual é bem definida.
54	modelos de estabelecimentos populacionais ao longo do tempo, possibilitando a definição de programas estratégicos de controle.
55	O estádio fenológico foi definido quando 50% + 1 das plantas apresentavam determinada característica de desenvolvimento.

Como é possível observar e conforme o esperado, a ER retorna fragmentos que contêm "defin", que podem se referir a flexões do verbo "definir" e também a palavras como "definição", "definitiva" ou até mesmo a desvios da língua padrão, como em "defini-se", por exemplo. Mesmo considerando esses ruídos, a etapa cumprida foi satisfatória, pois houve redução significativa de material que seria processado na etapa seguinte.

[98] Expressão regular é "uma composição de símbolos, caracteres com funções especiais, que, agrupadas entre si e com caracteres literais, formam uma sequência, uma expressão. Essa expressão é interpretada como uma regra, que indicará sucesso se uma entrada de dados qualquer casar com essa regra, ou seja, obedecer exatamente a todas as suas condições. (JARGAS, 2009, p.32)

Em um segundo momento, as ocorrências retornadas por meio da ER descrita anteriormente, para cada padrão verbal, foram agrupadas em um arquivo único, este foi submetido ao *software* de processamento de *corpus*, *Unitex*, para que selecionássemos as ocorrências que são formadas efetivamente por verbo, o que resultou na exclusão de fragmentos como 50, 51, 52 e 54.

O *Unitex* possui um dicionário da língua portuguesa do Brasil, acoplado ao seu sistema, portanto, é possível buscar pelo lema, entre colchetes angulares ("<LEMA>"), de forma que o sistema localiza no *corpus* todas as formas flexionadas do lema dado.

Ao final do processo, as seguintes frequências foram obtidas para os padrões verbais do tipo "analítico" no *corpus* de estudo (Tabela 16).

Tabela 16 – Ocorrência dos verbos "analíticos" no *corpus* de estudo

Lema	Busca por ER - número de ocorrências	Busca no Unitex - número de ocorrências
chamar	2877	1779
conceber	162	161
conhecer	7438	3397
denominarr	2373	2081
entender	1404	987
nomear	992	152

Fonte - Elaborado pela autora.

Na tabela 16, temos o lema na primeira coluna; a frequência da ER que representa o radical do lema na segunda coluna e, em seguida, na terceira coluna, são expostas as frequências dos lemas no *Unitex*.

Pelos números obtidos, observamos que o verbo que mais ocorre no *corpus* é "conhecer" e o que menos ocorre é "nomear". Também é possível concluir que houve uma diminuição significativa (cerca de 37%) do primeiro processamento do *corpus* por meio de ER para o processamento no *Unitex*.

Lembramos que, nesta etapa do trabalho, o objetivo foi recuperar as sentenças que apresentassem os verbos que estamos investigando, para que, a partir da leitura dos fragmentos em que eles ocorrem, pudéssemos analisar quais desses fragmentos constituiriam contextos definitórios.

7.5 ESTÁGIO 5 - IDENTIFICAÇÃO DOS CONTEXTOS DEFINITÓRIOS

A partir da relação de verbos potencialmente definitórios disponível, o passo seguinte foi analisar cada sentença com seus respectivos lemas, com o propósito de classificá-las entre contexto definitório (CD) e contexto não definitório (NCD).

Para tanto, foi adotado o protocolo descrito na subseção 7.5.1.

7.5.1 Protocolo de identificação de verbo definitório

Com os resultados de buscas gerados pelo *Unitex,* foi construído um *sub-corpus* referente a cada um dos lemas já explicitados na tabela 16. Primeiramente, foram excluídos dos sub-*corpora* alguns ruídos, como sentenças repetidas, sentenças em outras línguas (sentença 56) ou sentenças ininteligíveis (sentença 57).

56	{S} Define a national curriculum for radiology residents and test from it. {S}
57	{S} Para formar esse tipo de 1998) parece ser a de conceber a formação docente como um conjunto de vários saberes para responder às exigências específicas de situações concretas de ensino/aprendizagem. {S}

Em um primeiro momento, também cogitamos em desconsiderar casos de homonímia sintática em relação ao verbo, como o que ocorre com certa regularidade no *sub-corpus* do verbo "chamar", em que o substantivo feminino "chama" se confunde com o verbo "chamar" flexionado na 3ª pessoa do singular do presente do indicativo:

58	{S} A curvatura e arredondamento da borda da micropipeta foram obtidos com a chama de um isqueiro a gás. {S}.

Contudo, chegamos à conclusão de que tal ambiguidade linguística deve ser investigada*,* uma vez que o propósito da pesquisa é distinguir contextos definitórios dos não definitórios no âmbito do PLN. Dessa forma, não excluímos tais ocorrências do nosso rol de análise.

De forma geral, cada "verbo" apresenta-se dentro de uma sentença completa, com ressalva a algumas sentenças mais longas que, porventura, na compilação do *corpus* ou no seu processamento, sofreram algum tipo de problema.

Quando ocorreram esses casos, se o verbo e o seu contexto mais imediato estavam íntegros, como no exemplo 59, a sentença foi mantida no *corpus*.

59 {S} A normalidade aqui não é concebida como simples ausência de doença e de sofrimento, mas como resultado precário de estratégias defensivas elaboradas para resistir àquilo que, no trabalho, é desestabilizante e mesmo nocivo para o funcionamento psíquico e a saúde mental {S}

Devido à grande quantidade de dados que dispomos e de forma a facilitar o processo de identificação dos contextos definitórios, chegamos ao seguinte protocolo:

PI-01: Análise de agrupamento de sentenças que apresentam similaridade quanto à conjugação verbal, na seguinte sequência: particípio; gerúndio; infinitivo; 3ª pessoa do singular e do plural do indicativo e 1ª pessoa do singular e do plural; outras formas. Os agrupamentos configuraram-se desse modo, pois dessa forma foi possível analisar separadamente, para cada lema, a produtividade "definitória" de cada conjugação verbal; também foi possível observar como são construídos sintaticamente os contextos definitórios, uma vez que existe a possibilidade de a elaboração de um contexto definitório a partir de um verbo no gerúndio ser distinta de um verbo no particípio, por exemplo; ou ainda é possível examinar o que faz um contexto definitório com o verbo na 3ª. pessoa do singular se diferenciar desse mesmo verbo em um contexto não definitório.

PI-02: Desses mesmos agrupamentos, foram feitos subagrupamentos que compreendiam contexto similar à esquerda e, noutro momento, à direita. Seguem algumas sentenças (60 a 64), que apresentam o verbo "nomear" no infinitivo antecedido da preposição "para", a título de exemplo.

60 {S} Nos nossos casos, somente três (27,3%) lesões eram predominantemente sólidas, uma (9,1%) era cística e sete (63,6%) eram diferença no número de casos analisados nos dois estudos, e também devidas às diferenças metodológicas adotadas para nomear as lesões como sólidas, císticas e sólido-císticas. {S}

61 {S} Ao que parece, as espécies brasileiras do gênero Pomacea (ou Ampullaria) ainda são incompletamente conhecidas, carecendo também de melhores estudos a validade de diferentes designações, algumas das quais utilizadas para nomear uma mesma unidade taxonômica. {S}

62	{S} O Teste de Nomeação de Figuras por Escolha (TNF2.1-Escolha) avalia a habilidade de escolher palavras escritas para nomear figuras, e analisa processos quirêmicos, ortográficos e semânticos envolvidos. {S}
63	{S} Para nomear a figura de cabra, doze surdos escolheram a palavra{S}
64	{S} Posteriormente, era solicitado para nomear a comparação restante, relação BE, CE ou DE. {S}

PI-03: Para a identificação da sentença que contém um contexto definitório, o exame deve consistir no reconhecimento de uma expressão de "algo que é definido" (formada por um ou mais itens lexicais), do próprio verbo e da definição ou explicação (formada por um ou mais itens lexicais). A expressão de "algo que é definido" pode ser genericamente chamada "termo" (65), contudo, nesta análise, são considerados como tal palavras anafóricas (66), conceitos aparentemente não lexicalizados (67) e até "casa vazia" (68), conforme exemplificado nos contextos definitórios a seguir:

65	{S} Denomina-se de fator de bioconcentração (FBC) a relação entre a concentração do composto no tecido do organismo e na água na situação de equilíbrio, podendo este parâmetro também ser calculado por constantes cinéticas. {S}
66	{S} A diferença é que esta última é concebida como um sistema que realiza não só o armazenamento, mas também o tratamento das informações, característica mais fundamental e que requer mais investimento para a segurança. {S}
67	{S} Estes pesquisadores demonstraram que o acúmulo prolongado de frio (acima das reais necessidades das cultivares) caracteriza-se por antecipar mais a brotação do que a floração, e que gemas florais e vegetativas de uma mesma cultivar têm diferentes necessidades de calor durante a ecodormência. {S}
68	{S} É caracterizado por substituição do epitélio escamoso estratificado pelo metaplásico colunar especializado, contendo células caliciformes. {S}

Na primeira sentença, observa-se o termo complexo "fator de bioconcentração (FBC)". Na segunda sentença, a expressão "esta última" desempenha o papel de anáfora, retomando o termo proferido provavelmente na sentença anterior. Já na terceira sentença,

o sintagma “acúmulo prolongado de frio”, que não sabemos ao certo tratar-se de termo, é explicado. Por fim, a última sentença manifesta a definição, ao se iniciar por “é caracterizado por”, porém, o termo ao qual o verbo se refere está antes dessa sentença.

PI-04: Portanto, na análise é importante reconhecer as expressões referentes ao termo, ao verbo ou locução verbal e à definição ou explicação. Como o *corpus* é formado por artigos científicos, muitas vezes são usadas, com frequência, determinadas expressões que designam “estudo”, “trabalho”, “pesquisa”, entre outros, e estes fazem o papel de termo dentro de uma sentença aparentemente definitória, como na sentença 69:

69	{S} Este estudo se caracteriza como teórico-empírico em função do tipo de condução epistemológica e metodológica. {S}

Reconhece-se que, nessa sentença, a expressão “este estudo” se refere não a “estudo” em si, e nem a um termo explicitado na sentença anterior, mas sim ao trabalho apresentado no artigo. Se houver somente a substituição da expressão “este estudo” por qualquer outra expressão que designe um termo, a sentença poderá ser classificada como contexto definitório. Porém, é importante assinalar a sentença como contexto não-definitório, pois por meio desses “falsos positivos” tem-se acesso a expressões como “estudo” que podem servir em uma gramática de exclusão que irá eliminar sentenças desse tipo.

PI-05: Outro ponto relevante no protocolo de identificação dos contextos definitórios se refere ao foco no verbo que está sendo analisado. Há casos em que a sentença apresenta um contexto definitório, porém, a definição não é desencadeada pelo verbo, para o qual a sentença foi selecionada. Segue um exemplo (70) que consta no sub-*corpus* do verbo “caracterizar” e no sub-*corpus* do verbo “ser”:

70	{S} Os atributos sabor, textura e aparência externa, foram os que melhor caracterizaram o ponto de consumo e aceitabilidade do mamão, pelos provadores podendo, também, ser avaliados pela razão entre sólidos solúveis totais e acidez titulável, que é uma característica que reflete a qualidade sensorial de frutos(...). {S}

Nessa sentença, o verbo “caracterizar” não desempenha função de ligar um termo a uma definição. Entretanto, na mesma sentença, o verbo “ser” se manifesta

relacionando um termo (“razão entre sólidos solúveis totais e acidez titulável”) à sua definição (“característica que reflete a qualidade sensorial de frutos”).

Em casos como esses, a sentença não deve ser considerada como definitória na identificação de contexto definitório construído com o verbo que está sendo investigado (no caso, o verbo “caracterizar”). Em outros sub-*corpora*, pode ocorrer de o verbo que está sendo analisado se constituir como definitório, e a mesma sentença conter outro(s) verbo(s) que também é(são) definitório(s). Vejamos o exemplo 71, extraído do sub-*corpus* do verbo “caracterizar”:

71	{S} O agente etiológico – Pythium insidiosum é um oomiceto pertencente ao reino Stramenopila (ALEXOPOULOS et al., 1996) e caracteriza-se pela ausência de quitina na parede celular e ausência de esteróides na membrana plasmática. {S}

Como se nota, o “Pythium insidiosum” apresenta duas definições a partir dos verbos “ser” e “caracterizar”: “é um oomiceto pertencente ao reino Stramenopila (ALEXOPOULOS et al., 1996)” e “caracteriza-se pela ausência de quitina na parede celular e ausência de esteróides na membrana plasmática.” Em casos como esse, a sentença deve ser classificada normalmente como definitória, tanto no *sub-corpus* do verbo “caracterizar”, como no sub-*corpus* do verbo “ser”.

Outra possibilidade é o mesmo verbo ocorrer mais de uma vez na sentença. Quando isso acontecer, é feita a análise para cada um. Se pelo menos um verbo tiver o sentido definitório, a sentença deve ser classificada como contexto definitório. Segue um exemplo (72):

72	{S} O modelo de dados é mais bem entendido em um nível de abstração mais elevado do que o das tabelas, por meio de um esquema conceitual representando entidades, que podem ser entendidas como algo da realidade modelada em que se deseja manter informações no banco de dados, podendo representar tanto objetos concretos quanto objetos abstratos; {S}

PI-06: Hesitamos em classificar determinadas sentenças como contexto definitório ou como contexto não definitório. E de forma a não levar mais tempo na tarefa de identificação, quando houve esses casos, optamos por classificá-los como contexto definitório. Observe-se o exemplo 73.

73 {S} A grande abundância de peixes da família Ariidae na Baía de Sepetiba caracteriza a importância ecológica deste ambiente no ciclo de vida dessas espécies. {S}

Para finalizar esse tópico, é importante apontar que mesmo tendo chegado a esse conjunto de orientações na identificação de sentenças que constituem contexto definitório, a tarefa ainda assim é subjetiva, já que nem sempre é trivial sua classificação, seja pela dificuldade de reconhecer o termo ou a definição, seja pelo sentido empregado pelo verbo, seja pelo fato de as sentenças serem ambíguas, seja até mesmo pelo fato de, às vezes, o limite da sentença não ser suficiente para assegurar se a sentença é ou não um contexto definitório.

Na subseção 7.5.2, é descrita a constituição dos dois *corpora* de estudo, feitos para cada verbo: a) de contexto definitório (CD) e b) de contexto não-definitório (NCD).

7.5.2 Constituição do corpus de contexto definitório

A leitura, análise e classificação das sentenças culminaram na constituição de dois *corpora* para cada verbo apresentado no quadro 13. Os *corpora* foram denominados da seguinte forma:

Contexto Definitório (CD): *corpus* que inclui sentenças que apresentam um termo, uma definição ou explicação e o verbo que os relaciona. Esse *corpus* é a base do desenvolvimento de um conjunto de regras que posteriormente servirá para extrair contextos definitórios automaticamente. A seguir, são exibidas três sentenças (74, 75, 76) que fazem parte do *corpus* de CD.

74 {S} **A terceira versão do padrão,** nomeada **DICOM 3.0,** foi apresentada em 1993, quando foi criado o protocolo de comunicação para rede(6,7). {S}

75	{S} No presente trabalho, partindo de uma determinada concepção teórica que concebe **campo científico** como **um espaço de luta pelo monopólio da competência científica que é socialmente reconhecida a um agente determinado, ou seja, a capacidade técnica e o poder social de falar e intervir legitimamente em** {S}
76	{S} **A doença,** conhecida pelos produtores como **"sarna",** ocorre durante o período de chuvas, principalmente no final do mesmo, de março a maio, em propriedades em que as pastagens estão invadidas por grandes quantidades de **Froelichia humboldtiana** (Fig.1), conhecida **popularmente como ervanço.** {S}

No primeiro CD, o termo "DICOM 3.0" e a definição "A terceira versão do padrão" são relacionados pelo verbo "nomear" no particípio. No segundo CD, o termo "campo científico" é equivalente a "um espaço de luta pelo monopólio da competência científica" por meio do verbo "conceber". E no terceiro CD, o termo "sarna" e seu hiperônimo "doença" são ligados pelo verbo "conhecer", assim como o termo "Froelichia humboldtiana" e o seu sinônimo "ervanço".

Contexto Não Definitório (NCD): *corpus* que inclui sentenças que nitidamente não apresentam uma definição. Fazem parte desse *corpus*, sentenças que exibem o verbo com sentido não-definitório. Embora o foco da pesquisa esteja nas sentenças que se constituem como definitórias, um *corpus* de NCD pode auxiliar na compreensão dos elementos sintáticos ou semânticos que distinguem um NCD de um CD. Seguem três sentenças (77, 78, 79) que fazem parte desse *corpus*:

77	{S} Testar a hipótese que se estriba, fundamentalmente, na atividade anti- conceptiva do termofosfato magnesiano, específica em relação aos moluscos nomeados. {S}
78	{S} Foram observadas diferenças no GDM entre novilhas que conceberam ou que falharam apenas no grupo 18MP (0,621 vs 0,429kg/d), o que pode explicar o fato de as novilhas desse grupo não terem apresentado diferenças nos PIA entre prenhes e falhadas (302,1 vs 298,2kg). {S}
79	{S} Conhecer e aprender: sabedoria dos limites e desafios. {S}

Nos exemplos exibidos, extraídos do *corpus* de NCD, os verbos em destaque "nomeado", "conceberam" manifestam sentido aproximado de "eleito" no primeiro caso, e "gerar" no segundo caso. Já no último NCD, o verbo "conhecer" aparentemente se manifesta como um título.

Especificamente os *corpora* de CD apresentam apenas a sentença com o verbo. Não há anotação dos constituintes dos contextos definitórios (termo, definição e verbo), marcação alguma do arquivo em que o fragmento foi extraído, nem das sentenças adjacentes. Vale ressaltar que esses *corpora* já são produtos da pesquisa que poderão se empregados em investigações futuras sobre o tema.

Na seção 8, será feita uma breve discussão acerca da formação dos contextos definitórios referente aos verbos "nomear", "conceber", "chamar", "entender", "conhecer" e "denominar".

8. DESCRIÇÃO DOS VERBOS DEFINITÓRIOS

En la lengua, todo remite a diferencias, pero todo remite también a agrupamientos.

Ferdinand Saussure

Nesta seção, apresentamos a descrição do comportamento dos verbos constituintes de CD em termos quantitativos e qualitativos. Examinamos cada padrão verbal individualmente, de acordo com os seguintes aspectos: a) distribuição de sentenças que incluem um dos verbos investigados, em CD ou NCD; b) produtividade de cada estrutura sintática do verbo em combinação com o nexo (quando houver); c) distribuição de cada flexão verbal em relação às estruturas sintáticas observadas; d) verificação da distância entre o verbo e o seu nexo (quando houver) nos CDs encontrados e, por fim; v) identificação na sentença da posição em que o termo pode ocupar em relação ao verbo e ao nexo.

O propósito do exame pormenorizado de cada padrão verbal é, por meio dos dados obtidos, propor uma expressão regular, que subsidie a recuperação de contextos definitórios com base nos verbos investigados. Ressalte-se ainda que o foco dessa descrição consiste na análise do verbo e do nexo, portanto, neste momento, na sistematização dos dados, foram colocadas de lado informações como regras de exclusão e inclusão de determinados itens lexicais, que podem favorecer a otimização da recuperação dos contextos.

A seguir estão os seis verbos descritos: Nomear (8.1), Conceber (8.2), Chamar (8.3), Entender (8.4), Conhecer (8.5) e Denominar (8.6).

8.1 NOMEAR

"Nomear" foi o verbo que teve a menor frequência no *corpus* de estudo. Das 152 ocorrências do verbo que foram identificadas pelo *Unitex* no pré-processamento do *corpus*, 57 foram excluídas, por se tratar de sentenças incompreensíveis ou repetidas. Dessa forma, foi possível, de fato, a análise de 95 sentenças que continham o verbo "nomear", totalizando aproximadamente 3.240 *tokens*.

Na tabela 17, é exibida a distribuição das sentenças do *corpus* em CD e NCD.

Tabela 17 - Distribuição das sentenças do *corpus* "Nomear" em CD e NCD

CD	NCD	Total
16	79	95
17%	83%	100%

Fonte – Elaborado pela autora.

Como se observa, de acordo com a análise feita do *corpus*, na grande parte das vezes em que o verbo "nomear" ocorre (83%), não é com o intuito de apresentar uma definição ao leitor. É o que mostram as sentenças 80 e 81:

.

80 {S} Em 1996 foi nomeado coordenador do Projeto Radioastronomia - RA, no âmbito da Parceria INPE - UFSM. {S}

81 {S} Para nomear a figura de cadeira, vinte escolheram a palavra CAVEIRA. {S}

Na primeira sentença, "nomeado" apresenta a acepção de "designado", enquanto que, na segunda, "nomear" apresenta a acepção de "chamar". Embora seus significados se assemelhem entre si e com o próprio sentido do verbo quando num contexto definitório, podemos observar que as sentenças não apresentam uma explicação ou definição.

Em relação aos 16 CDs formados a partir do verbo "nomear", foi possível verificar que eles apresentam as seguintes estruturas sintáticas, independentemente da flexão do verbo (Tabela 18).

Tabela 18 – Estruturas sintáticas do padrão verbal definitório "Nomear"

ES1	**<Nomear> como X**	7	44%

	{S} O isolado de Marília foi nomeado como K5... {S}		
ES2	**<Nomear> de X** {S} Quando eram amplamente sólidas, foram nomeadas de sólidas... {S}	6	38%
ES3	**<Nomear> X** {S} A terceira versão do padrão, nomeada DICOM 3.0... {S}	2	13%
ES4	**<Nomear> com X** {S} ...nomeando este procedimento com fechamento primário retardado. {S}	1	6%
TOTAL		**16**	**100%**

Fonte – Elaborado pela autora.

Como é possível notar, as estruturas **<Nomear> como X e <Nomear> de X** foram as mais produtivas no *corpus* de estudo, somando 82%. Embora em menor quantidade, a **ES3**, que é formada pelo verbo sem nenhum nexo, era esperada que ocorresse, enquanto que a **ES4** nos surpreendeu.

Como se vê na Tabela 19, a flexão do verbo (**FL**) teve a seguinte distribuição de frequência no total e de acordo com as respectivas estruturas sintáticas (**ES**).

Tabela 19 – Distribuição da flexão verbal do padrão verbal definitório "Nomear"

		ES1	**ES2**	**ES3**	**ES4**	**TOTAL**	**TOTAL %**
FL5	Particípio	3	5	1	-	9	56%
FL6	Indicativo - Pretéritos, Presente e Futuro – 1 e 3ªP	3	1	1	-	5	31%
FL7	Gerúndio	1	-	-	1	2	13%
	TOTAL	**7**	**6**	**2**	**1**	**16**	
	TOTAL %	**44%**	**38%**	**12%**	**6%**		**100%**

Fonte – Elaborado pela autora.

As flexões relacionadas ao **FL5** e **FL6** apresentaram uma frequência mais significativa, 9 e 5 (56% e 31%) respectivamente, contra apenas 2 ocorrências do gerúndio (13%). Da mesma forma, das 6 vezes que a **ES2** ocorreu, 5 vezes foi com o verbo no particípio.

A distância máxima entre o verbo e o nexo foi de 2 palavras, como mostra a Tabela 20, referente à Janela de Palavras (**JP**) entre verbo e nexo.

Tabela 20 – Distância de palavras entre o verbo e o nexo do padrão verbal definitório "Nomear"

	ES1				ES2				ES4				TOTAL	TOTAL %
JP	0	1	2	3	0	1	2	3	0	1	2	3		
FL5	2	-	1	-	5	-	-	-	-	-	-	-	8	57%
FL6	1	-	2	-	1	-	-	-	-	-	-	-	4	29%
FL7	-	1	-	-	-	-	-	-	-	-	1	-	2	14%
TOTAL	3	1	3	-	6	-	-	-	-	-	1	-	14	
TOTAL %	21%	7%	21%	-	43%	-	-	-	-	-	7%	-		100%

Fonte – Elaborado pela autora.

É interessante observar que a **ES2,** formada por **<Nomear> de X**, sempre apresentou o nexo imediatamente após o verbo. As células que contêm "-" demonstram que não houve correspondência da estrutura sintática em questão quanto à flexão verbal e à janela de palavras entre o verbo e o nexo.

A posição do termo, em relação ao verbo e ao nexo, pode ocorrer conforme apresentado no Quadro 15.

Quadro 15 – Posição do termo do padrão verbal definitório "Nomear"

	ES1	ES2	ES3	ES4
FL5	DN	DN	DV	
FL6	DN / VN	DN	DV	
FL7	VN			DN

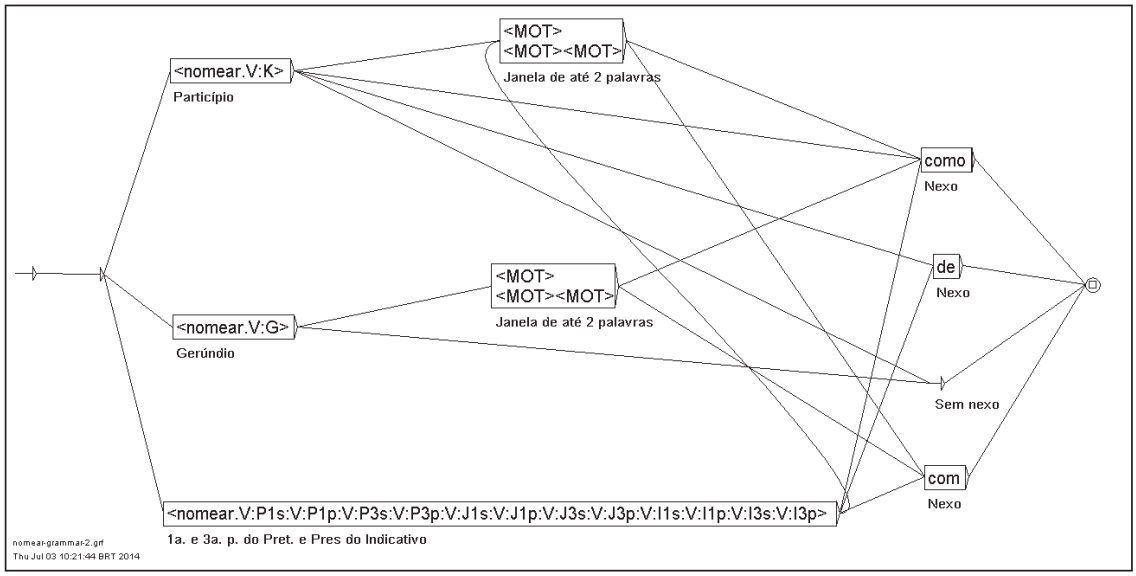
Fonte – Elaborado pela autora.

DN representa "direita do nexo", **DV** representa "direita do verbo" e **VN** representa "entre o verbo e o nexo". De acordo com a descrição, os termos podem se

posicionar à direita do verbo, quando for o caso da **ES2.** E, nos demais casos, podem se posicionar após o nexo (**DN**), ou entre o verbo e seu respectivo nexo (**VN**).

Com base no cruzamento dos dados observados quanto à estrutura sintática, flexão do verbo e distância entre o verbo e o nexo, foi elaborado o grafo no *Unitex,* apresentado na Figura 5, a título de observação.

Figura 5 – Grafo do padrão verbal definitório "Nomear"



Fonte – Elaborado pela autora.

Por meio desse grafo, é possível recuperar contextos definitórios como os apresentados em 82 e 83.

82	**ES1/FL6/DN/0**	{S} Dessa forma, Colle nomeia como sistema documental jornalístico todo sistema estruturado na forma de espaço de informação multidimensional, composto necessariamente por um conjunto de dados relacionados que permitem diversas formas de leitura e combinação. {S}
83	**ES1/FL5/DN/2**	{S} Os valores experimentais estão nomeados na legenda como TATe. {S}

Por outro lado, recupera-se também sentenças que não formam contexto definitório, como no exemplo 84.

84 {S} Testar a hipótese que se estriba, fundamentalmente, na atividade anti-conceptiva do termofosfato magnesiano, específica em relação aos moluscos nomeados. {S}

Casos como o exemplo 84 são possíveis de ocorrer, especialmente devido à estrutura **ES3**, que não apresenta o nexo após o verbo, elemento que restringe a busca. Contudo, pela descrição realizada, sabe-se que o termo vem imediatamente após o verbo nesse tipo de estrutura, de modo que esse problema pode ser minimizado, desde que o usuário conte com uma lista de termos, que poderá servir de entrada no sistema de busca por contexto definitório.

8.2 CONCEBER

Das 161 ocorrências do verbo "conceber", foram analisadas e classificadas 118. Sua distribuição ocorreu como consta na Tabela 21.

Tabela 21 - Distribuição das sentenças do *corpus* "Conceber" em CD e NCD

CD	NCD	Total
55	63	118
47%	53%	100%

Fonte – Elaborado pela autora.

De acordo com a classificação realizada, a distribuição de CDs e NCDs no *corpus* é bem equilibrada, com praticamente 50% cada. Ressaltamos que não houve dificuldades em classificar as sentenças em CD ou NCD, o que pode significar que a distinção do tipo de contexto em que aparece o verbo "conceber" é menos difícil do que para outros verbos.

As sentenças 85 e 86 são exemplos de NCDs.

85 {S} As novilhas que conceberam eram 18 dias mais velhas do que as que não conceberam na estação. {S}

86	{S}Por outro lado, o modo como são concebidos e implantados esses programas parece ser determinante importante do seu sucesso. {S}

A sentença 85 exemplifica o emprego de "conceber" com o significado de "dar à luz", enquanto que a sentença 86 exibe o verbo com o significado de "projetar".

Na Tabela 22, são expostas as estruturas sintáticas observadas no *corpus* de CD do verbo "conceber".

Tabela 22 - Estruturas sintáticas do padrão verbal definitório "Conceber"

ES1	**<Conceber> como X** {S}Destacamos que a comunicação pode ser concebida como um sistema de múltiplos canais...{S}	31	56%
ES2	**<Conceber> para X** {S}A operação evolucionária (EVOP), pouco conhecida, foi concebida para reduzir a variação em processos industriais. {S}	15	27%
ES3	**<Conceber> em|de|com X** {S}Esse software, concebido em linguagem C por intermédio do Turbo C (Borland), permite identificar, selecionar e subtrair...{S}	6	11%
ES4	**<Conceber> que X** {S}Esta abordagem concebe que a aprendizagem do adulto pode ser transformativa, ou seja, que pode gerar mudanças de consciência (Mezirow, 1991). {S}	1	2%
ES5	**<Conceber> por X** {S} A andragogia foi inicialmente concebida por Knowles, em 1975, como um modelo para a educação de adultos. {S}	2	4%
TOTAL		**55**	**100%**

Fonte – Elaborado pela autora.

São, no total, cinco possibilidades de arranjo sintático, das quais a primeira (**ES1**) se revela como a mais proeminente, haja vista a sua frequência de 56%. A segunda, a **ES2**, formada pelo verbo "conceber" seguido da preposição "para," teve uma frequência

significativa, de quase 30%. A **ES3** teve frequência total de 11%, e as duas últimas estruturas obtiveram frequência de apenas 2% (**ES4**) e 4% (**ES5**).

Na Tabela 23, é apresentada a frequência da flexão do verbo (**FL**) individualmente, no total e em relação com as respectivas estruturas sintáticas (**ES**).

Tabela 23 – Distribuição da flexão verbal do padrão verbal definitório "Conceber"

		ES1	ES2	ES3	ES4	ES5	TOTAL	TOTAL %
FL5	Particípio	19	15	6	-	2	42	76%
FL6	Indicativo - Pretéritos, Presente e Futuro – 1 e 3ªP	10	-	-	1	-	11	20%
FL7	Gerúndio	1	-	-		-	1	2%
FL8	Infinitivo	1	-	-		-	1	2%
	TOTAL	31	15	6	1	2	55	
	TOTAL %	44%	38%	12%	0%	6%		100%

Fonte – Elaborado pela autora.

A flexão verbal mais frequente no *corpus* de estudo foi a **FL5,** que é formada pelo particípio do verbo, com 76% do total. Em seguida, vem a **FL6** com 20%. As flexões formadas de gerúndio e infinitivo somam 4%. Ao observar a Tabela 23 por outro ponto de vista, é possível notar que os CDs a partir do verbo "conceber" se mostram mais produtivos quando o verbo está no particípio seguido do nexo "como" (**ES1 e FL5**) ou "para" (**ES2 e FL5**) e quando o verbo se realiza nas 1ª. e 3ª. pessoas no indicativo, seja nos pretéritos ou presente, seguido de "como" (**ES1 e FL6**). É válido destacar também que a **FL6** não apresentou ocorrência com a **ES2.**

Quanto à quantidade de palavras entre o verbo e o nexo, pôde-se notar que grande parte das sentenças apresenta o verbo seguido do nexo (cerca de 60%). As sentenças que apresentam uma janela entre o verbo e o nexo varia entre 1 e 2 palavras (cerca de 30%), e o intervalo de 2 a 7 (cerca de 7%). Nesse *corpus*, a sentença 87 apresentou o verbo a uma distância de 14 palavras do seu nexo "como".

87	{S}A idéia do TIPS foi inicialmente concebida por radiologistas da Universidade de Oregon, em Portland, EUA, no final dos anos 60, como uma extensão do procedimento de colangiografia transjugular(1,2). {S}

Embora essa sentença possa ser considerada como uma exceção, é interessante observar que esse tipo de informação pode ocorrer entre o verbo e o nexo. Nesse exemplo em específico, há o registro de "quem" (por radiologistas da Universidade de Oregon), "onde" (em Portland, EUA) e "quando" (no final dos anos 60) "a ideia do TIPS" foi "concebida".

A posição do termo na sentença, quanto ao verbo e ao nexo, de acordo com as **ESs** e **FLs,** pode ter a configuração apresentada no Quadro 16.

Quadro 16 - Posição do termo do padrão verbal definitório "Conceber"

	ES1	**ES2**	**ES3**	**ES4**	**ES5**
FL5	EV	EV	EV		EV
FL6	VN			DN	
FL7	VN				
FL8	VN				

Fonte – Elaborado pela autora.

Quando o verbo estiver no particípio, independentemente da **ES**, o termo se posicionará à esquerda do verbo, como exibido no exemplo 88.

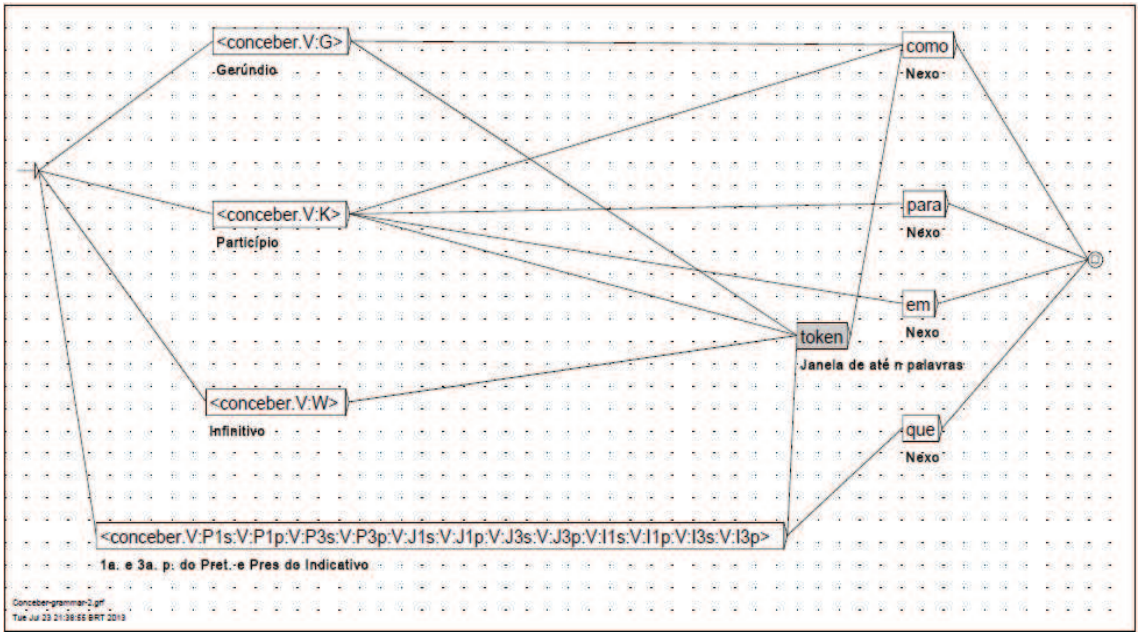
88 {S} Os riscos são concebidos como construções sociais, que se fazem por referência à identidade dos sujeitos... {S}

Já nos demais contextos definitórios relativos à **ES1 (FL6, FL7** e **FL8),** o termo se posiciona entre o verbo e o nexo, como se observa na sentença 89.

89 {S}De forma genérica, concebe-se a erosão como um processo de retirada...{S}

Para finalizar a descrição do verbo "conceber" em contextos definitórios, apresentamos na Figura 6 o grafo elaborado a partir dos dados observados.

Figura 6 - Grafo do padrão verbal definitório “Conceber”



Fonte – Elaborado pela autora.

Com esse grafo, foi possível recuperar sentenças como a apresentada no exemplo 90.

90	{S} Nos primeiros estágios de sua evolução, o Twin Block foi concebido como um aparelho removível simples, com bloco de mordida para colocar a mandíbula para frente. {S}

Destacamos que, a princípio, para esse padrão verbal, foi determinada uma distância máxima entre o verbo e o nexo de até 18 palavras. Isso significa que o grafo identificou contextos definitórios que atendiam a esse requisito, porém, do mesmo modo, o grafo recuperou raros NCDs, como o exibido na sentença 91.

91	{S} A segunda possibilidade seria considerar qualquer novilha cuja probabilidade de concepção fosse maior ou igual a 50% como passível de conceber, e aquela cuja probabilidade de gestação fosse menor que 50%, como passível de não conceber. {S}

Na subseção 8.3, é realizada a descrição do verbo “chamar”.

8.3 CHAMAR

Das 1.779 ocorrências de “chamar”, foram analisadas e classificadas 1.682, ou seja, 94% do total de ocorrências. Sua distribuição ocorreu conforme o exibido na Tabela 24.

Tabela 24 - Distribuição das sentenças do *corpus* “Chamar” em CD e NCD

CD	NCD	Total
903	779	1682
54%	46%	100%

Fonte – Elaborado pela autora.

Observa-se que a quantidade de CDs com “chamar” supera a quantidade de NCDs com uma diferença de 8%. Isso pode ser justificado devido ao fato de que foram consideradas também as ocorrências do particípio “chamar” precedido de artigo definido, que se confunde com o adjetivo, conforme os exemplos 92 e 93.

92	{S} Simultaneamente assalariado e representante do capital, ele integra a chamada nova classe média. {S}
93	{S} É o chamado "paradoxo polar" que confirma os achados para tocoferóis e ácido ascórbico como antioxidante. {S}

O emprego desse tipo de estrutura ocorreu 160 vezes no nosso *corpus*, o que representa cerca de 17% do total de CDs. Outro ponto interessante de mencionar em relação à formação desse tipo de construção diz respeito à localização da definição, pois ao contrário dos verbos já descritos, nesses casos, o contexto definitório pode abranger outras orações ou parágrafos manifestados antes ou depois do verbo “chamar”, incluindo outras relações semânticas importantes, haja vista que não há uma equação estabelecida como em “X é nomeado Y”.

Já as sentenças classificadas como NCDs se referem, sobretudo: a) ao verbo “chamar” + <DET> + “atenção”, totalizando cerca de 400 ocorrências; b) ao substantivo “chama(s)”, que ocorreu mais de 200 vezes. Observem-se os exemplos 94 e 95.

94	{S} A existência de outros trabalhadores sem qualificação técnica formal e regular, atuando no setor de coleta de exames, chama a atenção. {S}

95	{S} O sódio foi determinado por fotometria de chama, após ataque fluorídrico das amostras e o ferro bivalente por volumetria, após ataque nitroperclórico. {S}

Na sentença 94, é exibido o sintagma “chama a atenção”, que indubitavelmente não se refere a um contexto definitório, mas sim a uma expressão retórica que indica surpresa e relevância da informação relatada anteriormente: “A existência de outros trabalhadores sem qualificação técnica formal e regular, atuando no setor de coleta de exames”. Na sentença 95, há a questão da homonímia entre as formas do substantivo feminino “chama” com o verbo “chamar” na 3ª pessoa do singular do presente do indicativo.

Os contextos definitórios constituídos com o verbo “chamar” apresentam as estruturas sintáticas detalhadas na Tabela 25.

Tabela 25 - Estruturas sintáticas do padrão verbal definitório “Chamar”

ES1	<Chamar> de X {S} O cabo mensageiro de sustentação é, por sua vez, fixado aos postes, através de uma ferragem metálica chamada de braço suporte tipo "L" (SARDETO, 1999). {S}	434	48%
ES2	<Chamar> X|“X” {S}Neste caso, as constantes e sucessivas uniões dessas mulheres, chamadas "monogamia seriada", podem significar um esforço para manter no lar a figura do provedor (Bilac, 1995). {S}	469	52%
TOTAL		**903**	**100%**

Fonte – Elaborado pela autora.

As **ESs** detectadas demonstraram frequência bem equilibrada, 48% e 52% e, além disso, são **ESs** intercambiáveis, ou seja, a **ES1** pode substituir a **ES2** e vice-versa sem alteração de significado, demonstrando que a transitividade do verbo “chamar” com o significado de “ter por nome” pode ser direta **(ES2)** ou indireta **(ES1).**

Na Tabela 26, são exibidos os dados relativos às flexões do verbo “chamar” quanto às **ES1** e **ES2.**

Tabela 26 - Distribuição da flexão verbal do padrão verbal definitório “Chamar”

		ES1	ES2	TOTAL	TOTAL %
FL5	Particípio	335	449	784	87%
FL6	Indicativo - Pretéritos, Presente e Futuro – 1 e 3ªP	71	11	82	9%
FL7	Gerúndio	4	1	5	1%
FL7	Infinitivo	24	8	32	4%
	TOTAL	**434**	**469**	**903**	
	TOTAL %	**48%**	**52%**		**100%**

Fonte – Elaborado pela autora.

Ao analisar os valores absolutos, mais uma vez, percebe-se que a **FL** mais proeminente foi a do particípio, com 87% do total das ocorrências. Em segundo lugar, o indicativo com 9%. Em sequência, seguiram-se os verbos nas demais formas nominais, gerúndio (1%) e infinitivo (4%). Relacionando os dados referentes às **FLs** com as **ESs**, nota-se que o verbo no indicativo se mostra quase 7 vezes mais frequente na **ES1** do que na **ES2** e, em menor escala, o mesmo ocorre com as **FL7 e FL8.** Apenas no particípio a **ES2** é mais frequente do que a **ES1.**

Quanto à janela de palavras entre o verbo e o nexo, dá-se a configuração apresentada na Tabela 27.

Tabela 27 - Distância de palavras entre o verbo e o nexo do padrão verbal definitório “Chamar”

	ES1				TOTAL	TOTAL %
JP	0	1	2	3		
FL5	316	14	3	2	**335**	**77%**
FL6	65	3	3	-	**71**	**16%**
FL7	1	-	3	-	**4**	**1%**
FL8	19	5	-	-	**24**	**6%**
TOTAL	**401**	**22**	**9**	**2**	**434**	
TOTAL %	**92%**	**5%**	**2%**	**1%**		**100%**

Fonte – Elaborado pela autora.

Em 92% das ocorrências do verbo “chamar” em CDs, ele foi seguido diretamente do termo. Nos outros 8%, houve janela de 1 a no máximo 3 palavras.

A localização do termo em relação ao verbo e ao nexo (no caso da **ES1**) se mostrou com regularidade, como se nota no Quadro 17.

Quadro 17 - Posição do termo do padrão verbal definitório “Chamar”

	ES1	ES2
FL5	DN	DV
FL6	DN	DV
FL7	DN	DV
FL8	DN	DV

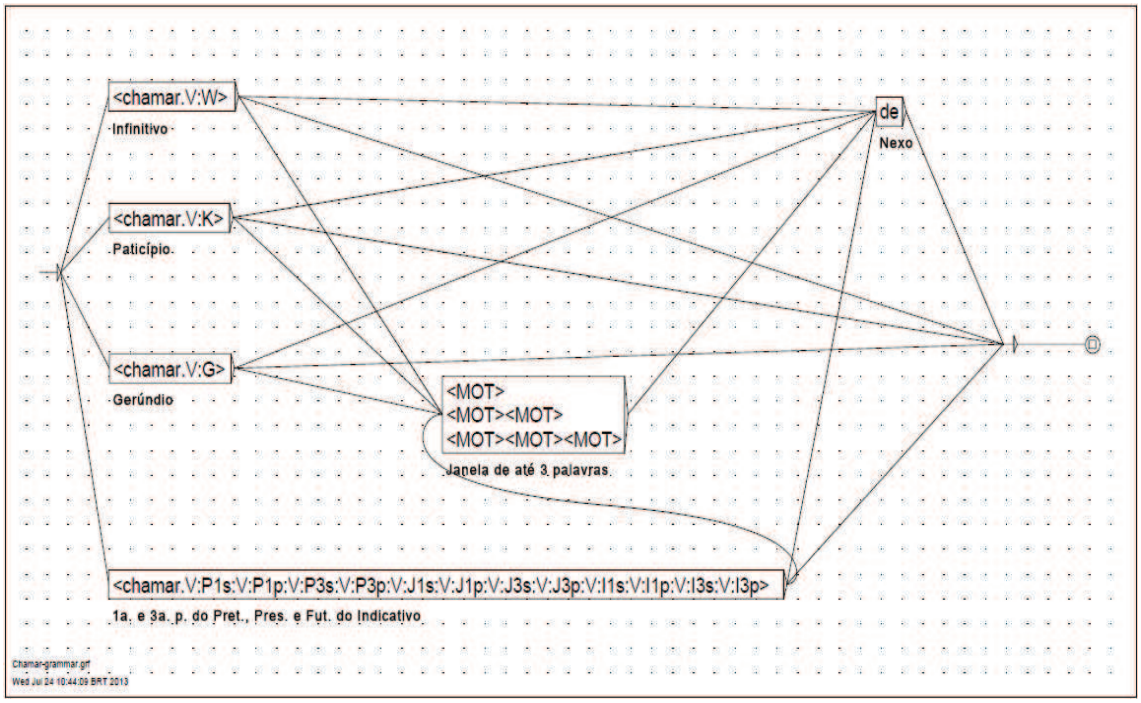
Fonte – Elaborado pela autora.

No caso de sentenças que apresentam a **ES1**, independentemente do tipo de **FL**, o termo se localiza à direita do nexo (**DN**). Nas **ES2**, o termo se localiza à direita do verbo (**DV**). Todavia, reconhece-se que à esquerda do verbo pode ser expresso um termo sinonímico, conforme o exemplo 96.

96	{S} A Sigatoka amarela, também chamada Cercosporiose da bananeira (Musa spp.)... {S}

Por fim, o grafo exibido na Figura 7 reúne os dados observados em relação ao comportamento do verbo “chamar” nos CDs.

Figura 7 – Grafo do padrão verbal definitório “Chamar”



Fonte – Elaborado pela autora.

O grafo permite identificar sentenças como o exemplo 97.

97	{S} O exocório seria formado pelo que chamou de exocório rijo, uma fina e resistente membrana de corionina que faz o revestimento externo do cório...{S}

Porém, esse grafo, sem contar com regras de exclusão ou com uma lista de termos, também identifica NCDs como o exemplo 98.

98	{S} Entre elas, a que mais nos chamou a atenção diz respeito à assistência oferecida à criança. {S}

8.4 ENTENDER

Das 987 ocorrências do verbo "entender", foram analisadas e classificadas 840, isto é, 85% do total de ocorrências. Sua classificação entre CDs e NCD pode ser observada na Tabela 28:

CD	NCD	Total
296	544	840
35%	65%	100%

Fonte – Elaborado pela autora.

Nota-se que a quantidade de NCDs é quase o dobro da quantidade de CDs. As sentenças 99 e 100 demonstram duas ocorrências de NCDs.

99	{S}Com a identificação dos problemas que abalam a humanidade percebe-se que estes não podem ser entendidos isoladamente, pois estão todos interligados e são interdependentes. {S}
100	{S} Entendendo a necessidade e a importância da disponibilização de modelos que permitam a geração de dados de precipitação para serem

	utilizados em diversas áreas, em particular em aplicações ligadas à agricultura e no dimensionamento de obras {S}

Quanto à estruturação sintática do padrão definitório "entender", encontramos cinco possibilidades, que são descritas na Tabela 29.

Tabela 28 - Estruturas sintáticas do padrão verbal definitório "Entender"

ES1	**<Entender> X|"X" como** {S} Após a colheita, o resfriamento rápido, entendido como a rápida remoção do calor de campo dos produtos, deve ser um dos primeiros passos a ser dado. {S}	209	71%
ES2	**<Entender> por X|"X"** {S} Entendemos por baixa qualidade da vida uma inserção social inadequada, reflexo do baixo salário e precárias condições de alimentação, moradia, transporte, vestuário e acesso à educação. {S}	47	16%
ES3	**<Entender> que X|"X"** {S} O tipo de relação entre os parceiros foi classificado em estável, entendendo que a relação estável é aquela que envolve, além de situações de casamento ou união consensual, relação afetivo-sexual com relações sexuais regulares. {S}	33	11%
ES4	**Por X|"X" <Entender>** {S}Por inteligência entendemos a capacidade de selecionar, comparar e combinar, com mais eficiência e rapidez, informações relevantes e irrelevantes; {S}	6	2%
ES5	**Como X|"X" <Entender>** {S} Como PCPI entende-se o período em que o controle da vegetação infestante deve ser realizado obrigatoriamente, situando-se entre os limites superiores do PAI e do PTPI. {S}	1	0%
TOTAL		**296**	**100%**

Fonte – Elaborado pela autora.

As duas primeiras **ESs** apresentam o verbo "entender" seguido das partículas "como" (**ES1**) e "por" (**ES2**). A primeira forma, **ES1**, é a que mais ocorreu no *corpus*,

com 71%. Já a segunda forma ocorreu em 16% das vezes. A **ES3** se refere ao verbo "entender" seguido da partícula "que" e representa 11% dos casos. As **ES4** e **ES5** apresentam as mesmas partículas das **ES1** e **ES2** ("como" e "por"), porém, estas se manifestaram antes do verbo "entender".

Na Tabela 29, são exibidas as frequências relativas a flexões do verbo "entender" quanto às **ESs** detectadas.

Tabela 29 - Distribuição da flexão verbal do padrão verbal definitório "Entender"

		ES1	ES2	ES3	ES4	ES5	TOTAL	TOTAL %
FL5	Particípio	170	11	-	-	-	181	61%
FL6	Indicativo - Pretéritos, Presente e Futuro – 1 e 3ªP	16	36	31	6	1	90	30%
FL7	Gerúndio	12	-	1	-	-	12	4%
FL8	Infinitivo	11	-	1	-	-	13	4%
	TOTAL	209	47	33	6	1	296	
	TOTAL %	44%	38%	12%		6%		100%

Fonte – Elaborado pela autora.

Assim como ocorreu com os verbos analisados nas subseções anteriores, o particípio é o tipo mais frequente, com 61%. Em seguida, o mais frequente é o **FL6**, com 30%. O gerúndio (**FL7**) e o infinitivo (**FL8**) apresentaram a mesma frequência, de 4% cada. Ao analisar as **ESs** em relação às **FLs**, observa-se que o padrão mais recorrente foi o "entendido como" com 170 vezes, em um total de 296 CDs.

Quanto à posição do termo em relação ao verbo e à partícula (nexo), tem-se o cenário apresentado no Quadro 18.

Quadro 18 - Posição do termo do padrão verbal definitório "Entender"

	ES1	ES2	ES3	ES4	ES5
FL5	EV/VN	EV			

FL6	VN/DN	DN	DN	NV	NV
FL7	VN		DN		
FL8	VN		DN		

Fonte – Elaborado pela autora.

É possível notar que, por exemplo, o termo se manifesta à esquerda do verbo (EV) quando ele está no particípio, como na sentença **ES1** da Tabela 29.

Quanto à quantidade de palavras entre o verbo e o nexo, o valor máximo encontrado foi de 3 palavras, considerando a combinação de **ES1** com **FL5**; e no máximo 1 palavras em padrões definitórios de: a) **ES3** com **FL6**, b) **ES2** com **FL6**, e c) **ES1** com **FL6** com o termo à direita do nexo (**DN**).

Na seção 8.1.5, o verbo "conhecer" é detalhado.

8.5 CONHECER

Das 3.397 ocorrências do verbo "conhecer", foram analisadas 3.076. Desse total, 78% foram classificadas como NCD e 22%, como CDs, como se observa na Tabela 31.

Tabela 30 - Distribuição das sentenças do *corpus* "Conhecer" em CD e NCD.

CD	NCD	Total
680	2396	3076
22%	78%	100%

Fonte – Elaborado pela autora.

As sentenças 101 e 102 apresentam o verbo "conhecer" como NCD.

101	{S} Portanto, em vista da necessidade de misturas de herbicidas para aumentar o espectro de controle e eficiência, torna-se primordial conhecer, também, os efeitos destas misturas sobre as culturas. {S}
102	{S} Assim, é fundamental conhecer o período de decomposição dos restos culturais de milho. {S}

Foram detectadas seis possibilidades de ESs, as quais são apresentadas na Tabela 32.

Tabela 31 - Estruturas sintáticas do padrão verbal definitório "Conhecer"

ES1	**<Conhecer> como X** {S} Tendo em vista estes aspectos, foi selecionada a espécie nativa Drimys brasiliensis Miers, conhecida como cataia. {S}	533	78%
ES2	**<Conhecer> por X** {S}Em todas as propriedades onde se observou a enfermidade havia a presença da árvore Ateleia glazioviana, popularmente conhecida por "timbó", "Maria preta", "cinamomo bravo" ou "amargo". {S}	127	19%
ES3	**<Conhecer> X** {S}O uso do adesivo biológico de fibrina é conhecido desde 1909, quando Bergel documentou o efeito hemostático do pó de fibrina. {S}	2	0%
ES4	**<Conhecer> que X** {S}É conhecido que a carbamazepina induz seu próprio metabolismo quando o tratamento é de longa duração, tendo, conseqüentemente, o aumento dos níveis séricos desta droga durante o tratamento(12). {S}	10	1%
ES5	**<Conhecer> com X** {S}A espécie Cryptocarya aschersoniana Mez, conhecida com o nome popular de canela-batalha, é uma importante espécie nativa, pertencente ao grupo ecológico das espécies clímax tolerantes à sombra (DAVIDE et al., 1995). {S}	2	0%
ES6	**Como é|são <Conhecer> X** {S} A lagarta-do-cartucho, como é conhecida popularmente, pode gerar grandes perdas na produção e na qualidade dos grãos, sendo controlada por meio da aplicação de inseticidas do grupo dos piretróides e fosforados (OMOTO et al., 2002). {S}	6	1%
TOTAL		**680**	**100%**

Fonte – Elaborado pela autora.

As **ESs** mais produtivas foram **ES1** e **ES2**, com 78% e 19%, respectivamente. As demais **ESs**, juntas, somam cerca de 3%. Observem-se os dados na Tabela 33.

Tabela 32 - Distribuição da flexão verbal do padrão verbal definitório "Conhecer"

		ES1	ES2	ES3	ES4	ES5	ES6	TOTAL	TOTAL %
FL5	Particípio	533	127	2	10	2	6	**680**	
	TOTAL %	**78%**	**19%**	**0%**	**1%**	**0%**	**1%**		**100%**

Fonte – Elaborado pela autora.

A única **FL** encontrada no *corpus,* que sugere que um fragmento é definitório, se manifesta somente no particípio.

Quanto à posição do termo, constata-se o que está apresentado no Quadro 19.

Quadro 19 - Posição do termo do padrão verbal definitório "Conhecer"

	ES1	ES2	ES3	ES4	ES5	ES6
FL5	DN/EV	DN/EV	EV	DN	DN/EV	DV/EN

Fonte – Elaborado pela autora.

Nota-se que, com exceção do **ES3** e **ES4**, os demais casos permitem que o termo se manifeste ora à direita do nexo (**DN**), ora à esquerda do verbo (**EV**) nas **ES1** e **ES2**; à esquerda do nexo (**EN**) e à direita do verbo (**DV**) na **ES6**, ou ainda à **DN** e à esquerda do verbo (**EV**) na **ES5**. Essa variedade de posição do termo em padrões definitórios de "conhecer" se deve ao fato de que o verbo é altamente produtivo em sentenças nas quais o autor apresenta uma definição por sinonímia, portanto, o termo pode se configurar tanto como *definiens* (definição), como *definiendum* (termo). As **ES1**, **ES2** e **ES5** da Tabela 32 são exemplos desse tipo de construção.

Em relação à quantidade de palavras entre o verbo e o nexo, apresentaram-se como valores máximos os expostos na Tabela 34.

Tabela 33 - Distância de palavras entre o verbo e o nexo do padrão verbal definitório "Conhecer"

JP	ES1	ES2	ES3	ES4	ES5	ES6
FL5	4	2	-	0	1	2

Fonte – Elaborado pela autora.

8.6 DENOMINAR

"Denominar" apresentou no total 2.081 ocorrências no *corpus*. Com a exclusão de ocorrências problemáticas, foram analisadas 1.850. Desse total, 99% foram classificadas como CD, conforme exibido na Tabela 35.

Tabela 34 - Distribuição das sentenças do *corpus* "Denominar" em CD e NCD

CD	NCD	Total

1831	19	1850
99%	1%	100%

Fonte – Elaborado pela autora.

Nota-se que, praticamente, todas as vezes que o verbo “denominar” ocorre, é possível a identificação: de como um conceito é chamado (103), ou do hiperônimo do termo (104), ou ainda de termo sinônimo (105).

103	{S}Denomina-se intensidade de precipitação máxima admissível o maior valor de precipitação que se pode aplicar em uma área sem provocar escoamento superficial. {S}
104	{S}Adotou-se então um procedimento que denominamos "estratégia do silêncio". {S}
105	{S}O pseudoperfilhamento, também denominado superbrotamento, constitui, sem dúvida, um dos maiores problemas na cultura do alho, principalmente nas cultivares que produzem o chamado "alho nobre". {S}

Já as sentenças 106 e 107 foram classificadas como NCDs.

106	{S}Se o volume de produção justificar o investimento em instalações para o beneficiamento de subprodutos, dentre as rações estudadas, a denominada para rãs é recomendada para o aproveitamento de fígado e descartes, e a aquela para trutas inicial, para aproveitamento de fígado e corpo gorduroso. {S}
107	{S}Identicamente, a taxa de câmbio também pode ser afetada por intervenções do Banco Central e pela classificação do risco soberano para a dívida denominada em moeda externa. {S}

Quanto às **ESs**, foram observados 6 possíveis arranjos, conforme exibidos na Tabela 36.

Tabela 35 - Estruturas sintáticas do padrão verbal definitório “Denominar”

ES1	**<Denominar> de X**	460	25%

	{S}Fitoplasmas são agentes causais de doenças de plantas, tendo sido descobertos em 1967 e denominados, na época, de organismos do tipo micoplasma ou MLOs (Doi Kirkpatrick, 1994). {S}		
ES2	**<Denominar> como X** {S}Foi convencionado denominar como galactomanano de D. mollis a fração de sementes sem casca retida na peneira de 420mm. 2.2- Caracterização das sementes casca e resíduos de hidróxido de sódio. {S}	38	2%
ES3	**<Denominar> X** {S}Uma operação agrícola, para ser efetiva no sistema de produção, deve ser executada no prazo agronomicamente ótimo, e a isso se denomina pontualidade da operação. {S}	1303	71%
ES4	**<Denominar> por X** {S}O conjunto dessas variáveis aleatórias foi denominado por IQP e constitui o vetor aleatório das variáveis climáticas. {S}	14	1%
ES5	**X (assim|como (ser)) <Denominar>** {S}A cirurgia ortognática é assim denominada por constituir-se de técnicas de osteotomias realizadas no sistema mastigatório com o objetivo de corrigir as discrepâncias relacionais maxilares e, por conseguinte, estabelecer o equilíbrio entre a face e o crânio. {S}	13	1%
ES6	**X utilizado|usado para <Denominar>** {S}Surgiu até um termo em inglês "critters", utilizado para denominar as pessoas envolvidas com a administração crítica (Fournier e Grey, 2000). {S}	3	0%
TOTAL		**1831**	**100%**

Fonte – Elaborado pela autora.

A **ES3** foi a mais recorrente no *corpus* (71%), seguida da **ES1** (25%). As demais estruturas totalizam 4%. Na Tabela 37, organizam-se as 6 **ESs** em relação ao comportamento flexional do verbo.

Tabela 36 - Estruturas sintáticas do padrão verbal definitório "Denominar"

								TOTAL	
		ES1	**ES2**	**ES3**	**ES4**	**ES5**	**ES6**	**TOTAL**	**%**
FL5	Particípio	416	29	1225	12	13	-	1695	93%
FL6	Indicativo	37	5	59	2	-	-	103	6%
FL7	Gerúndio	3	0	9	0	-	-	12	1%
FL8	Infinitivo	4	4	10	0	-	3	21	1%
	TOTAL	460	38	1303	14	13	3	**1831**	

TOTAL %	25%	2%	71%	1%	1%	0%	**100%**

Fonte – Elaborado pela autora.

A combinação mais produtiva foi do tipo **ES3** e **FL5** (com 66%), ou seja, com o verbo flexionado no particípio, seguido imediatamente do termo, como no exemplo 108.

108 {S}Lesões extracorticais de localização mais ventral, denominadas lesões profundas de malacia, não têm sido relatadas anteriormente na maioria dos trabalhos que descrevem a histopatologia do SNC de bovinos infectados espontânea ou experimentalmente por BHV-5 (Johnston et al. 1962, Barenfus et al. 1963, {S}

Quanto à posição do termo em relação ao verbo, notamos que ela se apresenta de modo estável, como se observa no Quadro 20.

Quadro 20 - Posição do termo do padrão verbal definitório "Denominar"

	ES1	ES2	ES3	ES4	ES5	ES6
FL5	DN	DN	DV	DN	EV	-
FL6	DN	DN	DV	DN	-	-
FL7	DN	-	DV	-	-	-
FL8	DN	DN	DV	-	-	EV

Fonte – Elaborado pela autora.

Como já explicitado anteriormente, o verbo "denominar" caracteriza-se por relacionar muitas vezes termos sinônimos. Dessa forma, embora no Quadro 20 tenha sido privilegiada certa posição do termo em relação ao verbo, isso não significa que é o único termo em sentenças desse tipo.

Em relação à quantidade de palavras entre o verbo e o nexo, apresentaram-se como valores máximos os expostos na Tabela 38.

Tabela 37 - Distância de palavras entre o verbo e o nexo do padrão verbal definitório "Conhecer"

JP	ES1	ES2	ES3	ES4	ES5	ES6
FL5	4	6	1	4	5	-
FL6	10	6	0	0	-	-
FL7	1	-	2	-	-	-
FL8	0	3	0	-	-	**2**

Fonte – Elaborado pela autora.

Assim como com os demais verbos, a grande maioria das sentenças são compostas pelo verbo seguido imediatamente do nexo, de forma que esses valores máximos correspondem a algumas poucas ocorrências. Além disso, algumas estruturas permitem que a definição se posicione entre o verbo e o nexo, como no exemplo 109, o que faz com que a distância entre eles seja maior.

109	{S}Denominamos o instrumento construído sob a forma de escala de Likert de "Escala de Atitude frente a AIDS" (EA-AIDS) - Anexo. {S}

De acordo com as análises realizadas, podemos observar que existem pontos em comum entre os seis verbos descritos, como a frequência do particípio e da proximidade entre o verbo e o nexo na formação de contextos definitórios. Por outro lado, chamou-nos a atenção as diferentes possibilidades de arranjos sintáticos.

A última seção (seção 9) dessa parte do trabalho é dedicada à apresentação dos recursos construídos a partir das análises feitas dos *corpora* de CD e NCD, voltados para a tarefa de extração de contextos definitórios.

9. RECURSOS PARA A EXTRAÇÃO DE CONTEXTO DEFINITÓRIO

A presente seção tem como objetivo apresentar o conhecimento linguístico construído a partir do estudo dos *corpora* de CD e NCD e da descrição dos padrões verbais exibida na seção 8, de forma a ser tratado computacionalmente.

Na subseção 9.1, nos dedicamos à formação da gramática de padrões verbais definitórios. Na subseção 9.2, exibimos a gramática de exclusão, e por fim, na subseção 9.3, expomos o conjunto de heurísticas para a classificação de contextos definitórios.

9.1 GRAMÁTICA DE PADRÕES VERBAIS DEFINITÓRIOS

Com o propósito de sistematizar o conhecimento linguístico que foi adquirido, desenvolvemos uma gramática local, ou seja, um conjunto de regras para cada padrão verbal definitório (subgramáticas). Apesar de os verbos que foram analisados apresentarem algumas estruturas semelhantes (em termos de tempo, pessoa, voz e de elemento que liga o termo à definição, o "nexo"), conforme exibido na seção 8, foi proposta neste trabalho uma investigação para cada verbo em particular, a fim de se verificar a ocorrência e produtividade dessas estruturas.

Cada subgramática é encabeçada por um conjunto de códigos, os quais explicitam as seguintes informações:

a) **ES_n:** refere-se à estrutura sintática do verbo, se ele está acompanhado de alguma partícula. Na gramática do verbo "chamar", há dois tipos de **ES**. A **ES1** que se refere ao verbo "chamar" seguido da preposição "de". A **ES2** que se refere a "chamar" seguido de nenhum nexo (elemento de ligação);

b) **FL_n:** diz respeito à flexão do verbo. Nessa gramática, **FL5** se refere ao particípio. **FL6** se refere às 1ª e 3ª pessoas do singular e plural no presente, futuro, pretérito perfeito e imperfeito do modo indicativo. **FL7** indica a forma no gerúndio; e **FL8**, a forma no infinitivo;

c) O terceiro código diz respeito à localização do termo que é definido em relação ao verbo ou ao nexo. Podem ser: **DV** (direita do verbo); **DN** (direita do nexo); **VN** (entre o verbo e o nexo).

d) Por fim, a quarta informação diz respeito à quantidade de *tokens* (cadeia de caracteres sem espaço) que pode ocorrer entre o verbo e o nexo (quando houver). Assim, **0** representa que não há nenhum *token* entre o verbo e o nexo; **1** representa que há um *token* entre eles, e assim por diante.

Ressaltamos que essas informações foram assim agrupadas em todas as gramáticas locais elaboradas, para que fosse possível fazer ajustes rapidamente e também com a finalidade de organizar o conhecimento obtido de modo simples.

A título de ilustração, segue a gramática local do verbo "chamar":

ES1-FL5-DN-0
CHAMADO DE TERMO
CHAMADOS DE TERMO
CHAMADA DE TERMO
CHAMADAS DE TERMO

ES1-FL5-DN-1
CHAMADO TOKEN DE TERMO
CHAMADOS TOKEN DE TERMO
CHAMADA TOKEN DE TERMO
CHAMADAS TOKEN DE TERMO

ES1-FL5-DN-2
CHAMADO TOKEN TOKEN DE TERMO
CHAMADOS TOKEN TOKEN DE TERMO
CHAMADA TOKEN TOKEN DE TERMO
CHAMADAS TOKEN TOKEN DE TERMO

ES1-FL5-DN-3
CHAMADO TOKEN TOKEN TOKEN DE TERMO
CHAMADOS TOKEN TOKEN TOKEN DE TERMO
CHAMADA TOKEN TOKEN TOKEN DE TERMO
CHAMADAS TOKEN TOKEN TOKEN DE TERMO

ES1-FL6-DN-0
CHAMO DE TERMO
CHAMA DE TERMO
CHAMAMOS DE TERMO
CHAMAM DE TERMO
CHAMAVA DE TERMO
CHAMÁVAMOS DE TERMO
CHAMAVAM DE TERMO
CHAMEI DE TERMO
CHAMOU DE TERMO

CHAMAMOS DE TERMO
CHAMARAM DE TERMO
CHAMAREI DE TERMO
CHAMARÁ DE TERMO
CHAMAREMOS DE TERMO
CHAMARÃO DE TERMO

ES1-FL6-DN-1
CHAMO TOKEN DE TERMO
CHAMA TOKEN DE TERMO
CHAMAMOS TOKEN DE TERMO
CHAMAM TOKEN DE TERMO
CHAMAVA TOKEN DE TERMO
CHAMÁVAMOS TOKEN DE TERMO
CHAMAVAM TOKEN DE TERMO
CHAMEI TOKEN DE TERMO
CHAMOU TOKEN DE TERMO
CHAMAMOS TOKEN DE TERMO
CHAMARAM TOKEN DE TERMO
CHAMAREI TOKEN DE TERMO
CHAMARÁ TOKEN DE TERMO
CHAMAREMOS TOKEN DE TERMO
CHAMARÃO TOKEN DE TERMO

ES1-FL6-DN-2
CHAMO TOKEN TOKEN DE TERMO
CHAMA TOKEN TOKEN DE TERMO
CHAMAMOS TOKEN TOKEN DE TERMO
CHAMAM TOKEN TOKEN DE TERMO
CHAMAVA TOKEN TOKEN DE TERMO
CHAMÁVAMOS TOKEN TOKEN DE TERMO
CHAMAVAM TOKEN TOKEN DE TERMO
CHAMEI TOKEN TOKEN DE TERMO
CHAMOU TOKEN TOKEN DE TERMO
CHAMAMOS TOKEN TOKEN DE TERMO
CHAMARAM TOKEN TOKEN DE TERMO
CHAMAREI TOKEN TOKEN DE TERMO
CHAMARÁ TOKEN TOKEN DE TERMO
CHAMAREMOS TOKEN TOKEN DE TERMO
CHAMARÃO TOKEN TOKEN DE TERMO

ES1-FL6-DN-3
CHAMO TOKEN TOKEN TOKEN DE TERMO
CHAMA TOKEN TOKEN TOKEN DE TERMO
CHAMAMOS TOKEN TOKEN TOKEN DE TERMO
CHAMAM TOKEN TOKEN TOKEN DE TERMO
CHAMAVA TOKEN TOKEN TOKEN DE TERMO
CHAMÁVAMOS TOKEN TOKEN TOKEN DE TERMO
CHAMAVAM TOKEN TOKEN TOKEN DE TERMO
CHAMEI TOKEN TOKEN TOKEN DE TERMO
CHAMOU TOKEN TOKEN TOKEN DE TERMO
CHAMAMOS TOKEN TOKEN TOKEN DE TERMO
CHAMARAM TOKEN TOKEN TOKEN DE TERMO
CHAMAREI TOKEN TOKEN TOKEN DE TERMO
CHAMARÁ TOKEN TOKEN TOKEN DE TERMO
CHAMAREMOS TOKEN TOKEN TOKEN DE TERMO
CHAMARÃO TOKEN TOKEN TOKEN DE TERMO

ES1-FL7-DN-0

CHAMANDO DE TERMO

ES1-FL8-DN-0
CHAMAR DE TERMO

ES1-FL7-DN-1
CHAMANDO TOKEN DE TERMO

ES1-FL8-DN-1
CHAMAR TOKEN DE TERMO

ES1-FL7-DN-2
CHAMANDO TOKEN TOKEN DE TERMO

ES1-FL8-DN-2
CHAMAR TOKEN TOKEN DE TERMO

ES1-FL7-DN-3
CHAMANDO TOKEN TOKEN TOKEN DE TERMO

ES1-FL8-DN-3
CHAMAR TOKEN TOKEN TOKEN DE TERMO

ES2-FL5-DV
CHAMADO TERMO
CHAMADOS TERMO
CHAMADA TERMO
CHAMADAS TERMO

ES2-FL6-DV
CHAMO TERMO
CHAMA TERMO
CHAMAMOS TERMO
CHAMAM TERMO
CHAMAVA TERMO
CHAMÁVAMOS TERMO
CHAMAVAM TERMO
CHAMEI TERMO
CHAMOU TERMO
CHAMAMOS TERMO
CHAMARAM TERMO
CHAMAREI TERMO
CHAMARÁ TERMO
CHAMAREMOS TERMO
CHAMARÃO TERMO

ES2-FL7-DV
CHAMANDO TERMO

ES2-FL8-DV
CHAMAR TERMO

9.2 GRAMÁTICA DE EXCLUSÃO

A gramática de exclusão tem como finalidade eliminar possíveis candidatos a CDs que apresentam comportamento sintático e de flexão verbal idênticos aos padrões verbais definitórios identificados nas gramáticas construídas, porém não são considerados como CD, como exemplificado em 80.

110 {S} O potássio foi quantificado por fotometria de chama, enquanto o cálcio, o magnésio e a dureza total o foram por titulometria (APHA, 1995; {S}

Nessa sentença, atribui-se à "chama" o sentido de "fogo", um caso de homonímia em relação ao verbo "chamar" na 3ª pessoa do singular do presente do indicativo. Sabendo que a subgramática **ES2-FL6-DV** representa, dentre outros, a expressão **CHAMA TERMO,** quando implementada no sistema de extração de contextos definitórios, sem apresentarmos um ou mais termos de entrada na busca pelo CD, se faz necessário o mapeamento de quais pistas linguísticas que circundam o verbo e o termo que poderiam indicar que o contexto se trata de um NCD.

Assim, de forma a tratar esse e outros fenômenos, realizamos um levantamento nas gramáticas locais, que permitiu identificar quais as pistas linguísticas que revelam falsos CDs. São elas (Quadro 21).

Quadro 21 – Gramática de exclusão

VERBO	SUBGRAMMAR	DV	DN	EV	EN	VN	NV
NOMEAR	ES3-FL5-DV ES3-FL6-DV	DE ACORDO COM <POR> . , (; ! ?					
CHAMAR	ES2-FL5-DV ES2-FL6-DV ES2-FL7-DV ES2-FL8-DV	(0-2) ATENÇÃO PARA . , (; ! ?					
	CHAMA			DE COM <POR> EM			
CONHECER	ALL	. , (; ! ?					
CONSIDERAR	ES3-FL5-EV-0 ES2-FL6-DV ES2-FL5-DV	. , (; ! ?					
CONCEBER	-						
DENOMINAR	ES3	. , (; ! ?					

Fonte – Elaborado pela autora.

Por meio desse mapeamento, será possível excluir do conjunto de candidatos a CDs sentenças que apresentam as sequências descritas no Quadro 21, como os exemplos 111, 112 e 113.

111 {S}Os corpos-de-prova da 2ª fase foram nomeados de acordo com as letras iniciais das variáveis estudadas. {S}

112 {S} Em relação aos motociclistas, chama a atenção o maior número de vítimas fatais observado nos meses de setembro, outubro e novembro. {S}

113 {S}O mecanismo de ação de A. glazioviana não é conhecido. {S}

Ressalte-se que optamos por não incluir na relação itens lexicais ou tipográficos que pudessem excluir sentenças que poderiam ser CD.

Além da proposta da gramática de exclusão, heurísticas para a construção de um classificador de CD também são apresentadas nesse trabalho, as quais são expostas na próxima subseção.

9.3 HEURÍSTICAS PARA UM CLASSIFICADOR DE CONTEXTOS DEFINITÓRIOS

A descrição quantitativa e qualitativa e, posteriormente, a avaliação dos padrões verbais definitório nos permitiu observar que há padrões que apresentam uma chance maior de serem classificados como CD, se comparados com outros padrões igualmente identificados e analisados nesse trabalho.

Dessa forma, com o propósito de otimizar a recuperação de contextos definitórios, exibindo um ranking com os candidatos a CDs em uma escala onde o mais provável candidato a CD seja exibido antes do menos provável candidato, é proposto um conjunto de heurísticas baseadas nos seguintes critérios:

a) Relevância do lema (H1): heurística que, a partir do total em porcentagem dos verbos que constituíram o *corpus* de estudo, considera o quanto o lema foi produtivo como CD. Assim, quanto mais um determinado lema foi classificado como constituinte de um CD, maior é o seu valor no ranking. Tal valor foi estipulado de forma simplificada: valor 1 atribuído para ocorrências entre 0% a 33%; valor 2 atribuído para 34% a 67% e valor 3 atribuído para ocorrências entre 68% a 100%.

Quadro 22 – Heurística 1 – Relevância do lema

H1		
Lema	**Ocorrências em %**	**Valor**
denominar	99%	3
chamar	54%	2
conceber	47%	2
entender	35%	2
conhecer	22%	1
nomear	17%	1

Fonte – Elaborado pela autora.

b) Relevância da subgramática (H2): heurística que considera a produtividade de uma subgramática em relação às outras subgramáticas de mesmo lema. Os valores considerados foram os seguintes: entre 0 a 10% é atribuído valor

1, entre 11 a 20% é atribuído valor 2, e assim sucessivamente até a margem 91 a 100% com o valor máximo de 10. Dessa forma, por exemplo, para a gramática dos verbos "nomear" e "conceber", teríamos os valores apresentados nos Quadros 23 e 24, respectivamente.

Quadro 23 – Heurística 2 - Relevância da subgramática (nomear)

H2 - Nomear		
Subgramática	**Ocorrências em %**	**Valor**
ES1-FL5	19%	2
ES1-FL6	19%	2
ES1-FL7	6%	1
ES2-FL5	31%	4
ES2-FL6	6%	1
ES3-FL5	6%	1
ES3-FL6	6%	1
ES4-FL7	6%	1

Fonte – Elaborado pela autora.

Quadro 24 – Heurística 2 - Relevância da subgramática (conceber)

H2 - Conceber		
Subgramática	**Ocorrências em %**	**Valor**
ES1-FL5	35%	4
ES1-FL6	18%	2
ES1-FL7	2%	1
ES1-FL8	2%	1
ES2-FL5	27%	3
ES3-FL5	11%	2
ES4-FL6	2%	1
ES5-FL5	4%	1

Fonte – Elaborado pela autora.

Ressalte-se que nessa heurística, foi desprezada a informação de algumas subgramáticas acerca da quantidade de *tokens* entre o verbo e o nexo, pois por se tratar de um dado relevante para o classificador, será apresentada uma heurística própria para o fenômeno.

c) Avaliação do verbo (H3): heurística que atribui valor para um lema, de acordo com sua taxa de precisão alcançada na avaliação. Assim como na H2, entre 0 a 10% é atribuído valor 1, entre 11 a 20% é atribuído valor 2, e assim sucessivamente até a margem 91 a 100% com o valor máximo de 10. Essa heurística é apresentada no Quadro 25.

Quadro 25 – Heurística 3 – Avaliação do verbo

H3		
Lema	**Ocorrências em %**	**Valor**
denominar	99%	3
chamar	61%	7
conceber	67%	7
entender	52%	6
conhecer	44%	5
nomear	64%	7

Fonte – Elaborado pela autora.

d) Localização da sentença no texto (H4): heurística que leva em conta a posição do padrão definitório no texto acadêmico. Como as seções Introdução e Metodologia são as que mais apresentam CDs, a elas são atribuídos os valores mais altos, como apresentado no Quadro 26.

Quadro 26 – Heurística 4 – Localização da sentença no texto

H4	
Seção do Texto	**Valor**
Resumo	1
Introdução	5
Metodologia	4
Discussão	1
Conclusão	1

Fonte – Elaborado pela autora.

e) Marcadores (H5): heurística que privilegia sentenças que, além do padrão definitório, apresentem determinados marcadores (substantivos, conectores e os tipográficos) que reforçam o caráter de CD do fragmento. O Quadro 27 traz alguns exemplos.

Quadro 27 – Heurística 5 - Marcadores

Marcador	Exemplo	Valor total
termo	{S} No entanto, tem-se verificado um aumento significativo na incidência de desordens fisiológicas conhecidas como "internal breakdown", termo usado para representar uma ou mais desordens fisiológicas caracterizadas pelo amadurecimento prematuro e desuniforme do mesocarpo da manga. {S}	2
palavra	{S} Anthrax (palavra originada do grego que significa carvão) é uma doença infecciosa aguda, também conhecida como carbúnculo por causar lesão cutânea negra(8). {S}	3
nome	{S} O seu limite oriental é representado por faixa de caráter transicional, conhecida pelo nome genérico de agreste e englobando vários aspectos florísticos de transição entre a floresta tropical atlântica e a caatinga propriamente dita. {S}	1
nomenclatura	{S} Segundo a classificação botânica de Swingle (Swingle e Reece, 1967), o limão Cravo denomina-se C. reticulata var. austera, nomenclatura que se refere, evidentemente, a uma tangerina. {S}	2
denominação	Assim, essa formação é também conhecida pela denominação de floresta caducifólia não espinhosa, embora para alguns essa denominação não seja apropriada, uma vez que não passa de vegetação de "arvoredo" com árvores disseminadas (Foury 29, 1968). {S}	4
isto é	{S} O desenvolvimento de uma RNA consiste em determinar sua arquitetura, isto é, os números de camadas e de neurônios em cada camada, bem como ajustar os seus parâmetros livres, fase esta conhecida como treinamento. {S}	1
ou seja	{S} A aprendizagem gerencial foi, por muito tempo, concebida sob essa ótica, ou seja, como um fenômeno que ocorria única e exclusivamente pela educação e pelo desenvolvimento gerencial, cuja ênfase está, respectivamente, na teoria e na prática (Fox, 1997). {S}	2
()	{S} O Trifoliata é conhecido como um porta-enxerto que induz frutos menores, de altos Brix e ratio, boa coloração e maturação tardia (Stuchi et al., 1996). {S}	1
“ ”	{S} Por sua vez, o resíduo basal, também chamado "coração" ou "palmito caulinar", é a porção do estipe imediatamente abaixo do meristema apical, que é bastante tenra para ser consumida "in natura". {S}	2

Fonte – Elaborado pela autora.

Na H5, é atribuído valor 1 a cada ocorrência de um marcador na sentença. Dessa forma, por exemplo, a sentença 1 possui valor 2, pois apresenta o marcador termo e aspas; a sentença 2 apresenta valor 3 por apresentar três marcadores: dois parênteses e o substantivo palavra, e assim por diante. Ressalte-se que essa relação não é fechada, sendo possível, portanto, incluir outros tipos de marcadores.

f) Ocorrências de outros verbos na sentença (H6): heurística caracterizada pela presença de dois ou mais verbos investigados no trabalho (nomear, conceber, chamar, entender, conhecer, denominar) na mesma sentença, como no nas sentenças 114 e 115.

114 {S} O pseudoperfilhamento, também denominado superbrotamento, constitui, sem dúvida, um dos maiores problemas na cultura do alho, principalmente nas cultivares que produzem o chamado "alho nobre". {S}

115 A primeira, conhecida também como andragogia, é uma abordagem para a aprendizagem na fase adulta pertencente ao paradigma humanista da educação, que concebe o processo de aprendizagem como um ato pessoal para realização do potencial humano (Merrian e Caffarella, 1991). {S}

Na sentença 114, coocorrem "denominar" e "chamar", enquanto que na sentença 115, estão presentes "conhecer" e "conceber". Essas pistas conferem às sentenças a confirmação de se constituírem, de fato, CDs. Assim, nessa heurística, é estabelecido valor 1 para cada sentença que apresentar dois ou mais verbos do conjunto descrito.

g) Tamanho da sentença (H7): heurística que considera que quanto maior for a sentença, melhor será seu CD, pois apresentará mais informação semântica, conforme os exemplos 116 e 117.

116	{S} O Indicador de Salubridade Ambiental, aqui definido como ISA/JP, foi concebido para servir como um instrumento eficaz na busca da salubridade, uma vez que aponta de forma sintética e eficiente as medidas que devem ser implementadas a fim de se obter melhorias na qualidade de vida, abrangendo os aspectos {S}
117	{S} O ensino é concebido com a função de transmitir informações. {S}

Observa-se que entre as duas sentenças, a 116 apresenta mais informações semânticas do que a 117, o que contribui para um melhor entendimento do termo. Portanto, no classificador, a sentença 116 deve ser exibida antes da sentença 117.

h) Distância entre o verbo e o nexo (H8): heurística que pondera a quantidade de *tokens* entre o verbo e o nexo (janela de palavras). Quanto menor for a janela de palavras entre eles, maior é a chance de o contexto apresentar uma definição ou explicação, conforme os exemplos 118 e 119.

118	A enfermidade é conhecida como "doença do peito inchado" em razão do edema subcutâneo de declive que ocorre como resultado da insuficiência cardíaca congestiva. {S}
119	{S}A soja verde já era conhecida de longa data e referida como "soja verde fresca" (fresh green soybeans) {S}

Nessa heurística, no caso dessas duas sentenças, a 118 deve ter melhor colocação no ranking em comparação com a sentença 119. Uma vez que foi observado que a grande maioria dos nexos ocorrem imediatamente após o verbo, essa heurística visa, sobretudo, a "penalizar" as sentenças que são falsos positivos, isto é, aquelas que apresentam aparentemente o verbo e seu nexo, quando, na verdade, não são. Um exemplo é a própria sentença 118, na qual se manifesta um "como" depois de 13 palavras, mas que não desse refere à palavra "conhecida".

Na seção 10, são apresentados as discussões e os resultados do trabalho.

PARTE 4

DISCUSSÃO, RESULTADOS E CONSIDERAÇÕES FINAIS

10. DISCUSSÃO E RESULTADOS DA PESQUISA

Se quiser evitar a arbitrariedade nas suas deduções, o estudioso dos problemas da linguagem terá que ascender da observação dos fatos da fala para concluir sobre os fatos da língua

Maria Teresa Camargo Biderman

Nesta seção, fazemos uma breve discussão e apresentamos os resultados sobre os principais aspectos da trajetória da pesquisa. Para tanto, dividimos a seção em quatro partes: metodologia (10.1), descrição (10.2), protótipo (10.3) e avaliação (10.4).

10.1 METODOLOGIA

A metodologia empregada quanto à organização e manipulação do *corpus* se mostrou eficaz. Porém, houve dificuldade inicialmente de estabelecer a quantidade ideal de textos e de ocorrências de verbos que deveria ser analisada. Sabíamos que poucos textos e, consequentemente, poucas ocorrências de parte significativa dos verbos poderiam invalidar a análise e a futura implementação; por outro lado, muitas ocorrências despenderiam muito tempo no processamento dos textos, na classificação da sentença, na análise propriamente dita e, dependendo da estrutura sintática ou da flexão do verbo, não seria tão relevante ter um maior número assim. Dessa forma, optamos por trabalhar inicialmente com os verbos que apresentassem incidência mais baixa (como "nomear," com 152 ocorrências) para depois analisar verbos mais frequentes (como "chamar," com 1.779 ocorrências). Essa decisão permitiu testar e aprimorar incrementalmente a metodologia, o uso de ferramentas e a sistemática da análise dos dados.

De fato, quanto maior a quantidade de dados de análise, mais tempo, concentração e cuidado deve ter o analista. Entretanto, criar um método de pesquisa a ser seguido e automatizar o que for possível tornam o trabalho mais exequível do que se realizado sem um protocolo e de forma completamente manual. Assim, o protocolo elaborado para a identificação de contextos definitórios foi útil para que pudéssemos criar procedimentos quanto a sua análise. Evidentemente, embora tenhamos planejado os passos de como identificar os contextos definitórios, esta, ainda assim, é uma tarefa subjetiva.

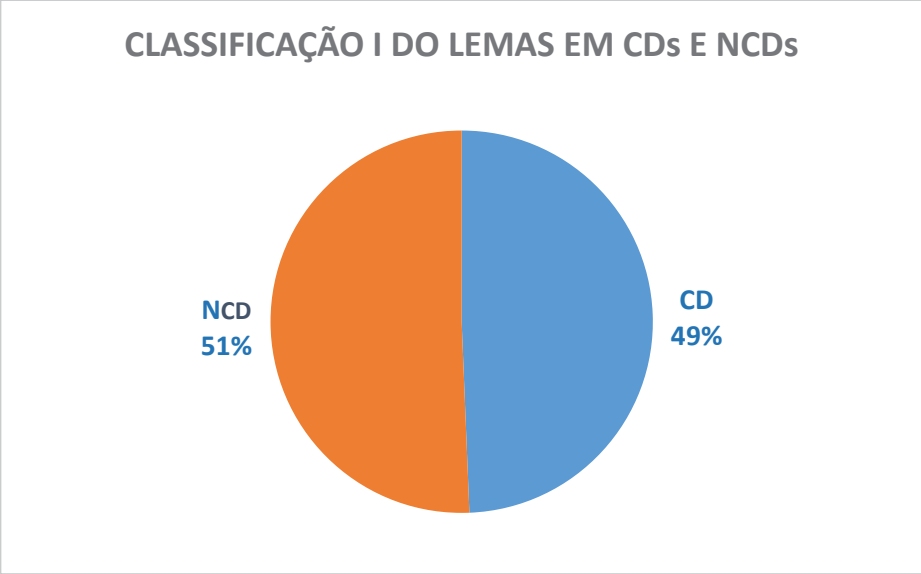
Entretanto, podemos dizer que a subjetividade se limita à classificação das sentenças entre contexto definitório e contexto não definitório. A descrição dos padrões verbais definitórios que foi realizada na seção 8 apresenta as evidências na superfície do texto. Assim, as informações fornecidas sobre os padrões definitórios verbais nos permitem aprender sobre seus vários aspectos, tais como: qual é a forma verbal mais produtiva para definir em um *corpus* acadêmico? Qual é a chance de um verbo "x" apontar para um contexto definitório? Existe uma localização prototípica de um verbo definitório em uma sentença? Essas e outras perguntas podem ser respondidas com o apoio dos dados gerados e com o suporte dos *corpora* de contextos definitórios criados para a pesquisa.

10.2 DESCRIÇÃO

A descrição dos padrões verbais definitórios com base em um *corpus* de artigos científicos nos permitiu concluir que:

a) Na metade das vezes em que os lemas analisados ocorreram no *corpus* de estudo, eles não foram usados para manifestar um contexto definitório. No Gráfico 1 são exibidos os valores percentuais totais de CDs e NCDs encontrados no *corpus*.

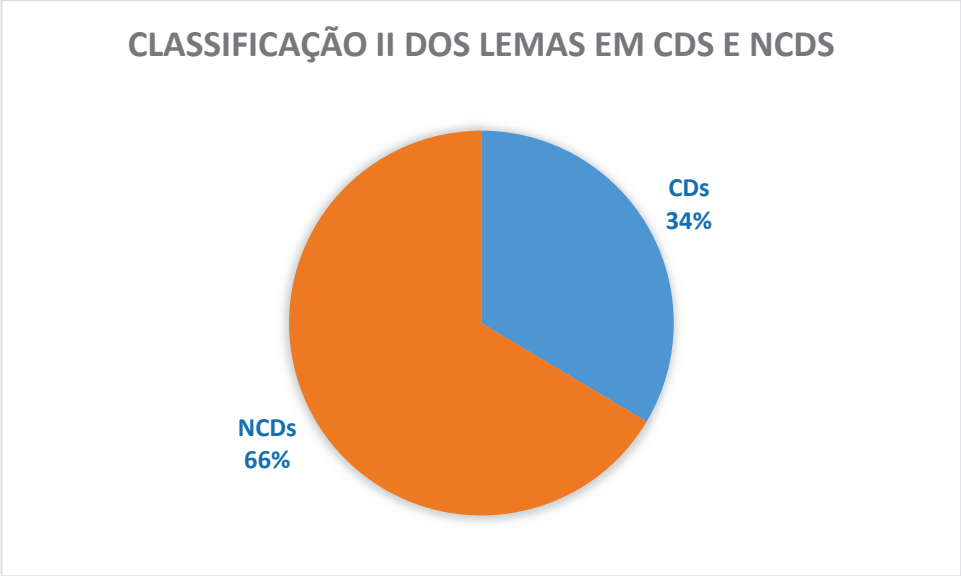
Gráfico 1 – Classificação I dos lemas em CDs e NCDs.



Fonte – Elaborado pela autora.

De modo geral, 49% das vezes em que os lemas ocorrem é para constituir contexto definitório. Contudo, se considerarmos a exclusão do verbo "denominar" que foi o mais produtivo na formação de CDs (com 99%), a produtividade dos lemas como constituintes de CDs teria uma redução de 15%, conforme exibido no Gráfico 2.

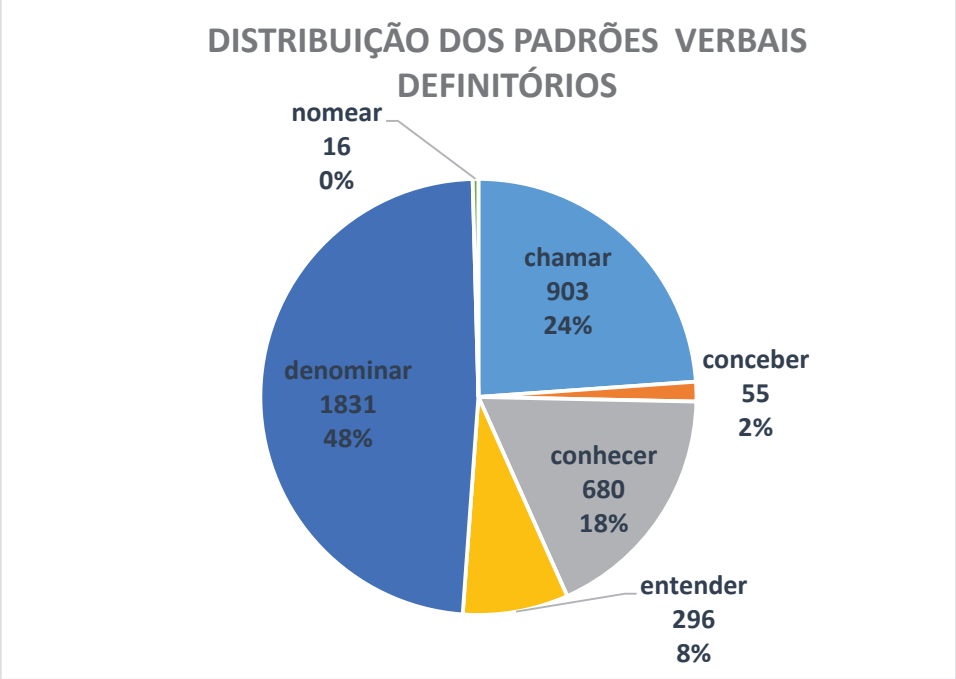
Gráfico 2 – Classificação II dos lemas em CDs e NCDs.



Fonte – Elaborado pela autora.

b) Do conjunto total dos CDs analisados (3.781 CDs), a distribuição dos padrões verbais definitórios ocorreu conforme apresentado no Gráfico 3.

Gráfico 3 – Distribuição dos padrões verbais definitórios

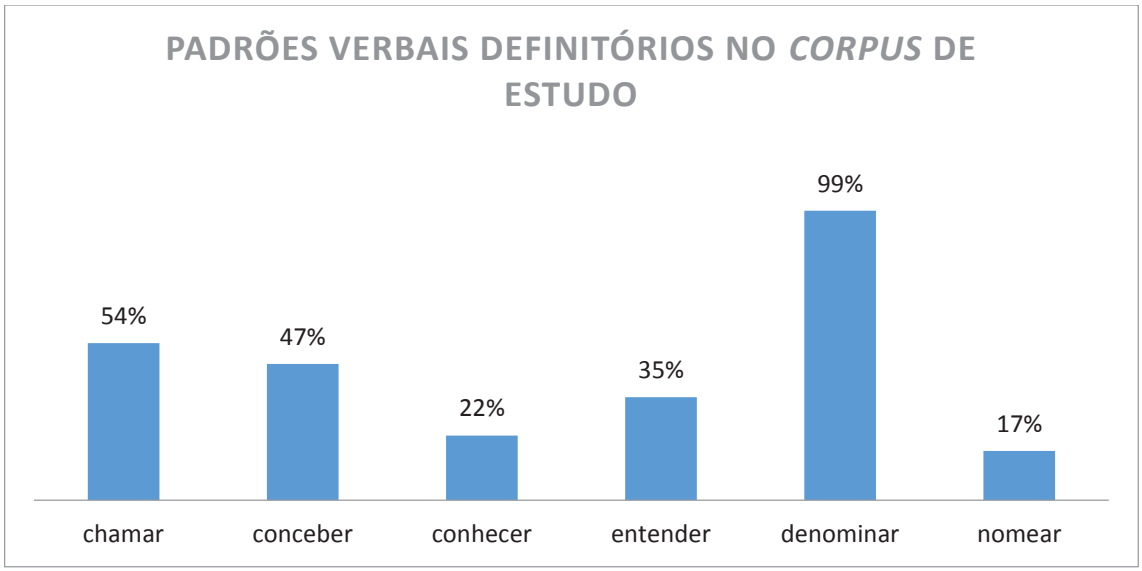


Fonte – Elaborado pela autora.

c) Os verbos "denominar", "chamar" e "conhecer" e suas flexões são ocorrências muito produtivas na língua, se comparadas à quantidade de ocorrências dos demais lemas do grupo (ver Tabela 16). O gráfico 3 sugere que esses três verbos também demonstram ser profícuos como constituintes de CD. No outro extremo, "conceber" e "nomear" tiveram pouca relevância, se considerarmos que se manifestaram apenas 2% e 0%, respectivamente, na soma total dos CDs detectados.

d) Considerando a porcentagem dos lemas que indicam um CD, tem-se o seguinte panorama, exibido no Gráfico 4.

Gráfico 4 - Padrões verbais definitórios no *corpus* de estudo



Fonte – Elaborado pela autora.

Apesar de "conceber" apresentar poucas ocorrências (2%) de CD no *corpus* de estudo (ver Gráfico 3), no Gráfico 4, nota-se especialmente que, de todas as vezes que o lema ocorreu no *corpus* de estudo, em quase metade das vezes (47%), ele foi classificado como CD. Dessa forma, é possível afirmar que, embora ele não seja um verbo muito produtivo na língua, "conceber" possui um índice relativamente alto de se manifestar em um CD.

e) Considerando o total de 3.781 CDs identificados no *corpus* de estudo, a partir dos seis verbos analisados, e a quantidade de 5.732 artigos científicos que o constituem, a razão entre CDs por artigos científicos é de 0,66 contextos definitórios por artigo científico. Considerando que foram analisados poucos verbos, em análises futuras de outros padrões verbais definitórios, possivelmente esse quociente irá aumentar.

f) Há CDs mais prototípicos, como os que são formados por particípio ("chamado", por exemplo), seguido de partículas "como" ou "de" ou ainda seguido do próprio termo. Os CDs formados por particípio representam 92%. Os padrões verbais definitórios no modo indicativo apresentam 5% do total de CDs examinados. Por último, seguem os CDs formados pelos verbos no gerúndio e infinitivo, com 1% e 2%, respectivamente.

Como é possível observar nessa breve discussão, a descrição realizada permite ter acesso aos dados organizados e cruzá-los de forma a testar algumas hipóteses. A partir daí, podemos tirar algumas conclusões, pelo menos dentro do nosso escopo de pesquisa.

Na subseção 10.3, é apresentado o protótipo criado com base nas gramáticas locais construídas.

10.3 PROTÓTIPO

Foi criado um protótipo de um Extrator de contextos definitórios[99] em desktop. O sistema foi desenvolvido em Java, com o suporte da biblioteca *java.regex* que possibilita a implementação das gramáticas locais por meio de expressões regulares em seu código.

O sistema permite que a entrada seja: i) um termo, ii) uma lista de termos, iii) sem um termo específico, iv) uma expressão regular. Além disso, é necessário fazer o *upload* do arquivo que o usuário deseja que seja processado.

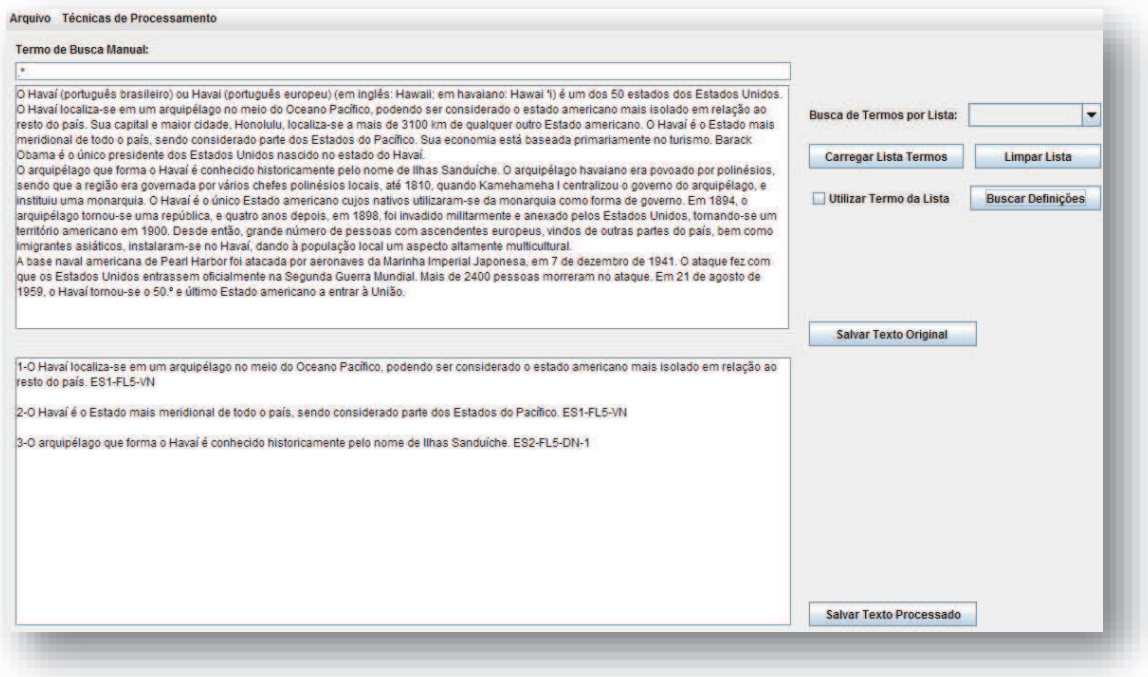
A primeira parte da execução do programa consiste em fragmentar o texto em sentenças. Em seguida, as sentenças que apresentam correspondência com uma das gramáticas de contextos definitórios são exibidas na tela do programa, junto com o respectivo código da subgramática correspondente.

Como o sistema é um protótipo, ainda se faz necessário agregar a ele uma série de funcionalidades como: a) *upload* de mais de um arquivo por vez; b) visualização de sentenças acima e abaixo da sentença identificada como contexto definitório, a critério do usuário; c) *hiperlink* na sentença para visualização da mesma no seu contexto maior, d) implementação dos recursos descritos nas seções 9.2 e 9.3.

Na Figura 8, apresentamos a tela do programa.

[99] O sistema foi desenvolvido pelos informatas Guilherme Ribeiro Olivatto e Thiago Ribeiro, como Trabalho de Conclusão do Curso Tecnologia em Análise e Desenvolvimento de Sistemas, no âmbito do Instituto Federal de São Paulo (IFSP), campus São Carlos, SP.

Figura 8 – Tela do protótipo “Extrator de contextos definitórios”.



Fonte – *Print screen* da tela do “Extrator de contexto definitórios”

No campo “termo de busca manual”, foi utilizada o curinga “.*” que significa qualquer caractere, com exceção da quebra de linha (“.”) entre 0 e ilimitadas vezes (*), o que significa que é possível retornar qualquer sequência de caracteres na posição do termo nas gramáticas elaboradas. Na *text area* superior, o texto entrada sobre o Havaí[100], extraído da Wikipédia, foi usado como exemplo no programa. Na *text area* abaixo, os candidatos a contextos definitórios são exibidos, seguidos de códigos de suas respectivas subgramáticas.

Reproduzimos, a seguir, as sentenças 120, 121 e 122 extraídas pelo programa.

120 {S} O Havaí localiza-se em um arquipélago no meio do Oceano Pacífico, podendo ser **considerado** o estado americano mais isolado em relação ao resto do país. {S}

121 {S} O Havaí é o Estado mais meridional de todo o país, sendo **considerado** parte dos Estados do Pacífico. {S}

122 {S} O arquipélago que forma o Havaí é **conhecido** historicamente pelo nome de *Ilhas Sanduíche* ("Sandwich Islands"). {S}

[100] <pt.wikipedia.org/wiki/Hava%C3%AD>. Acesso em fev. 2014.

Apesar das restrições do protótipo, as quais poderão ser trabalhadas posteriormente, foi possível utilizá-lo na avaliação das regras dos padrões definitórios construídas, a qual é apresentada na subseção 10.4.

10.4 AVALIAÇÃO

Para a avaliação, foram utilizados dois *corpora*. O primeiro se refere ao *corpus* do domínio "Sensoriamento remoto," desenvolvido pela Embrapa Informática Agropecuária. O *corpus* é composto por 316 artigos científicos, com um total de 860 mil *tokens*, aproximadamente. Como não conseguimos extrair quantidade suficiente de dados para a avaliação, foi imprescindível complementar com um segundo conjunto de textos, o qual se constitui como o *corpus* de artigos científicos de variadas áreas, do Banco do Português que não foi utilizado como *corpus* de estudo. Ele é composto por 1.911 arquivos, com 7,2 milhões de *tokens* aproximadamente.

Ao final da organização do *corpus* de avaliação, temos as seguintes quantidades de ocorrências dos lemas para a realização da avaliação (Tabela 39).

Tabela 38 – Quantidade de lemas usada na avaliação.

Lema	Ocorrências
Chamar	235
Conceber	49
Conhecer	161
Denominar	150
Entender	240
Nomear	18
Total	**853**

Fonte – Elaborado pela autora.

Três juízes participaram do processo de avaliação, os quais foram escolhidos devido à experiência com o trabalho terminológico. A tarefa deles consistiu na classificação dos fragmentos que contêm um dos seis verbos expressos na Tabela 39 em:

D: contexto definitório;

N: contexto não definitório;

B: fragmentos incompletos ou confusos;

S: incerteza se o fragmento é contexto definitório ou não definitório.

Além dessa classificação, os juízes foram orientados a classificar como D quando o fragmento apresentava um termo e uma explicação ou definição sobre ele, sendo que o verbo em destaque fazia ligação entre eles. O termo podia ser um termo simples (ex. *tarifa*), termo complexo (ex. *métricas de paisagem*), uma anáfora (ex. *essa máquina*), ou um conceito não lexicalizado (ex. *a água destinada à irrigação*, *abastecimento público ou navegação*).

O processo de classificação dos fragmentos teve duração aproximada de 7 horas e os três juízes avaliaram o mesmo conjunto dos 853 fragmentos e seus respectivos lemas. Para efeito dessa avaliação, o fragmento foi considerado como contexto definitório se ao menos um juiz o tivesse classificado com tal.

Com o propósito de avaliar o grau de concordância entre os três juízes na classificação das sentenças, foi empregado o coeficiente *Kappa* (COHEN, 1960) pois é considerado uma medida tradicional de concordância em PLN (HEARST, 1997). Este coeficiente tem como valor máximo 1 para K, onde este valor representa concordância perfeita entre os juízes; por outro lado, quanto mais próximo de 0 estiver o valor de K, mais este sugere que o grau de concordância entre os juízes se deve ao acaso. Valores negativos (até -1) refletem grau de concordância inferior ao esperado pelo acaso (FONSECA, SILVA, SILVA et al., 2007). O *Kappa* é expresso na seguinte fórmula:

$$K = \frac{P(a) - P(e)}{1 - P(e)}$$

Onde: P(a) é o total de concordância observada; P(e) é o total de acordo por coincidência.

Em nossa análise, os coeficientes *Kappa* obtidos para cada verbo se encontram na Tabela 40.

Tabela 39 – Coeficientes *Kappa*

LEMA	KAPPA-01	KAPPA-02
Chamar	0,606	0,781
Conceber	0,403	0,605
Conhecer	0,792	0,867
Denominar	0,280	0,433
Entender	0,639	0,79
Nomear	0,541	0,533
Média	**0,543**	**0,668**

Fonte – Elaborado pela autora.

Na aplicação do coeficiente *Kappa-01* foram consideradas três categorias: 1) N, 2) D e 3) S (S e B), estas últimas foram consideradas como uma única categoria, pois ambas indicam que a sentença não foi classificada de fato, demonstrando incerteza do juiz, falta de elemento na sentença ou confusão de elementos na sentença. Já no *Kappa-02*, foram consideradas apenas as sentenças categorizadas como N ou D, sendo as demais desprezadas nessa análise.

Embora não haja um valor específico a partir do qual se deva julgar o valor do *Kappa* como adequado, encontram-se na literatura algumas sugestões que orientam essa decisão. Por exemplo, Fleiss (1981) sugere que um valor menor do que 0.40 pode indicar uma concordância pobre; de 0.40 - 0.75, a concordância pode ser considerada de satisfatória a boa; e um valor maior de 0.75 é excelente.

Com esses parâmetros, podemos afirmar que a concordância geral na tarefa de classificação das sentenças como contexto definitório ou não definitório foi satisfatória (0,543 e 0,668), o que demonstra que as instruções para a tarefa foram claras, porém, esses dados também revelam que, mesmo com as devidas orientações, a prática de identificação de uma sentença como contexto definitório, muitas vezes, não é consensual.

Em seguida, os 853 fragmentos foram processados pelo protótipo Extrator de contextos definitórios, o qual identificou, segundo as regras criadas, quais deles eram contextos definitórios.

A fim de comparar a classificação dos juízes com a do protótipo, foram utilizadas as métricas de precisão (P) e cobertura (C). Neste trabalho, a precisão se refere à quantidade de contextos recuperados que são realmente definitórios; e a cobertura diz respeito à quantidade de contextos definitórios que são recuperados. As métricas são expressas nas seguintes equações:

$$C = \frac{vp}{vp + fn} \qquad P = \frac{vp}{vp + fp}$$

Onde:

***vn* (verdadeiro negativo):** fragmento não classificado como D pelos juízes e classificado como N pelo extrator. O *vn* é quando o extrator acerta na classificação da sentença em NCD.

***vp* (verdadeiro positivo):** fragmento classificado como D pelos juízes e classificado como D pelo extrator. O *vp* é quando o extrator acerta na classificação da sentença em CD.

***fn* (falso negativo):** fragmento classificado como D pelos juízes e classificado como N pelo extrator. O *fn* é quando o extrator erra na classificação da sentença em NCD.

***fp* (falso positivo):** fragmento não classificado como D pelos juízes e classificado como D pelo extrator. O *vn* é quando o extrator erra na classificação da sentença em CD.

O positivo, nesse caso, é quando o sistema classifica a sentença como CD; e negativo, quando o sistema a classifica como NCD. O verdadeiro se refere a quando o sistema acerta, e falso se refere a quando o sistema erra. Segue a Tabela 41 que sintetiza a descrição de cada categoria.

Tabela 40 – Classificação dos fragmentos na avaliação.

	VP	FP	VN	FN
EXTRATOR	D	D	N	N
JUIZES	D	≠D	≠D	D

Fonte – Elaborado pela autora.

Com a aplicação das métricas, chegamos aos resultados expressos na Tabela 42.

Tabela 41 – Avaliação das gramáticas de padrões definitórios.

Lema	VP	FP	VN	FN	P (%)	C (%)
Chamar	134	86	12	3	**61%**	**98%**
Conceber	28	14	4	3	**67%**	**90%**
Conhecer	52	65	44	0	**44%**	**100%**
Denominar	147	0	0	3	**98%**	**100%**
Entender	61	57	101	21	**52%**	**74%**

Nomear	7	4	6	1	**64%**	**88%**
Total	**316**	**327**	**287**	**32**	**64%**	**92%**

Fonte – Elaborado pela autora.

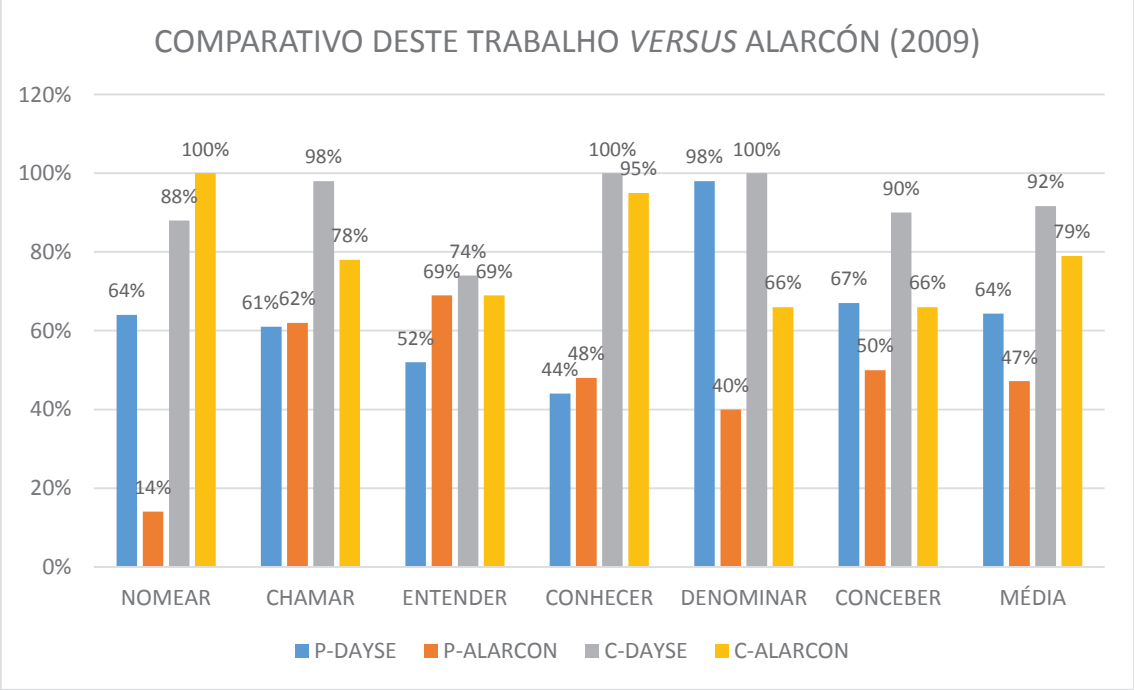
A avaliação realizada exibe o percentual de precisão e cobertura individualmente para os seis lemas investigados. A precisão alcançou um resultado mediano. Os melhores índices foram quanto aos verbos "conceber" (67%), "nomear" (64%) e "denominar" (98%), e o pior índice foi do lema "conhecer", com 44%.

Podemos notar que a cobertura atingiu um excelente resultado quanto aos lemas "denominar" (100%), "chamar" (98%) e "conhecer" (100%). A cobertura mais baixa foi do lema "entender", com 74%.

Considerando o total dos lemas, a avaliação obtida foi de 64% de precisão e 92% de cobertura. Esses resultados sugerem que, hipoteticamente, de 100 CDs identificados pelo extrator, 64 são CD realmente, e 8 CDs não foram identificados pelo extrator.

Ao comparar os resultados da avaliação com o trabalho de Alarcón (2009) (subseção 6.2.6), temos os dados expressos no Gráfico 5.

Gráfico 5 – Comparativo de avaliação I.

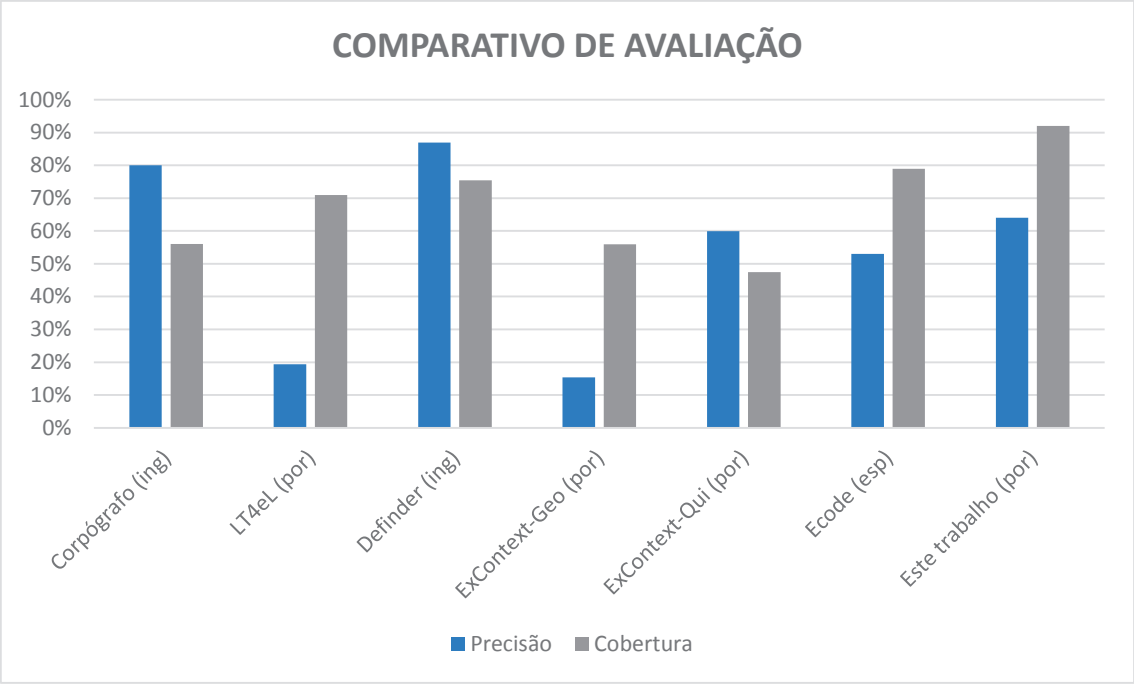


Fonte – Elaborado pela autora.

Nota-se que na avaliação das gramáticas dos CDs, o trabalho de Alarcón (2009) foi superior quanto à precisão (P) para os verbos "chamar" (62%), "entender" (69%) e "conhecer" (48%). Na outra metade dos verbos, este trabalhou apresentou melhor desempenho. Em relação à cobertura (C), nosso trabalho apresentou avaliação mais satisfatória do que a avaliação de Alarcón (2009), com exceção do verbo "nomear", com 100% contra 88%. O último conjunto de barras do gráfico diz respeito à média geral, e dele podemos extrair que a nossa gramática poderá apresentar um bom desempenho, se comparado com os resultados da pesquisa de Alarcón (2009).

Ao comparar ainda com alguns trabalhos descritos na subseção 6.2, tem-se o cenário expresso no Gráfico 6.

Gráfico 6 – Comparativo de avaliação II.



Fonte – Elaborado pela autora.

De acordo com os valores obtidos nas avaliações, as gramáticas desenvolvidas nesta pesquisa tiveram um maior êxito quanto à cobertura em comparação com os demais trabalhos. Em relação à precisão, as gramáticas para a língua inglesa do *Corpógrafo* e do *Definder* atingiram uma melhor avaliação do que a proposta aqui. Contudo, obtivemos uma melhor avaliação ao comparar nossos resultados com os resultados dos trabalhos voltados para a língua portuguesa. Observa-se que nossos resultados ainda podem ser

melhorados, considerando que, na avaliação, não foi considerada a gramática de exclusão apresentada na subseção 9.2.

Na próxima e última seção (11), são apresentadas as considerações finais da pesquisa.

11. CONSIDERAÇÕES FINAIS

> *Qual é, enfim, a utilidade da Linguística? Bem poucas pessoas têm a respeito ideias claras: não cabe fixá-las aqui. Mas é evidente, por exemplo, que as questões linguísticas interessam a todos – historiadores, filólogos etc. – que tenham de manejar textos (...) Seria inadmissível que seu estudo se tornasse exclusivo de alguns especialistas; de fato, toda a gente dela se ocupa pouco ou muito (...)*
>
> **Ferdinand Saussure**

Por fim, apresentamos as considerações finais do trabalho. Em primeiro lugar, na subseção 11.1, abordamos o tópico "Considerações gerais da pesquisa" com o objetivo de retomar os principais pontos do percurso da investigação. Em seguida, na subseção 11.2, asseveramos a originalidade da pesquisa realizada, tanto quanto à sua metodologia, quanto à análise gerada na descrição dos contextos definitório em português. Nas subseções 11.3, 11.4 e 11.5, trazemos possíveis contribuições do trabalho para a Terminologia, PLN e Linguística, respectivamente. Concluímos o texto, com algumas sugestões de trabalhos futuros na subseção 11.6.

11.1 CONSIDERAÇÕES GERAIS DA PESQUISA

Este trabalho teve como objetivos gerais: 1) investigar padrões de contextos definitórios presentes em *corpora* de especialidades em língua portuguesa; 2) proporcionar conhecimento linguístico que possa ser formalizado computacionalmente a fim de integrar um sistema de extração semiautomática de candidatos a contextos definitórios; 3) e avaliar os resultados gerados.

Para alcançar esses objetivos, foram executadas as seguintes tarefas:

a) apresentação da área de Terminologia e seus objetos que estão relacionados à etapa da redação da definição terminológica;
b) apresentação do que é contexto definitório nos estudos terminológicos, na Linguística Aplicada e na nossa investigação;
c) revisão bibliográfica de trabalhos que tiveram como escopo a identificação e a extração semi(automática) de contextos definitórios;
d) eleição do *corpus* e dos padrões definitórios para a análise;

e) criação de um protocolo para a identificação dos CDs;
f) compilação de *corpora* de CD e NCD para cada um dos seis verbos analisados;
g) descrição dos padrões verbais definitórios, com base na estrutura sintática e flexão do verbo, na localização do termo que é definido em relação ao verbo e ao nexo e na quantidade de *tokens* que pode correr entre o verbo e o nexo;
h) construção e organização das gramáticas de padrões verbais definitórios, da gramática de exclusão e das heurísticas para um classificador de CDs.
i) avaliação dos resultados gerados, considerando as métricas de precisão e cobertura.

Embora não tenhamos investigado todos os verbos do conjunto pensado inicialmente (29 no total), por meio dessas tarefas descritas, foi possível cumprir nossos três objetivos.

11.2 ORIGINALIDADE DO TRABALHO

A originalidade do trabalho reside no fato de que a abordagem na descrição quantitativa e qualitativa de verbos que se apresentam em CD é inédita em língua portuguesa, considerando a granularidade dos dados descritos. Prova disso é a quantidade de regras geradas para as gramáticas locais a partir da observação dos padrões verbais definitórios.

Além do objetivo de descrever o comportamento dos padrões verbais definitórios para fins de implementação, tínhamos como questão de pesquisa avaliar em que medida regras mais precisas quanto à formação dos CDs gerariam melhores resultados na sua extração semi(automática). Os resultados da avalição, embora sejam referentes apenas a seis verbos, parecem indicar uma resposta afirmativa a essa questão de pesquisa, haja vista que somente com a descrição do comportamento verbal, das partículas e da quantidade de *tokens* entre eles, foi possível obter 64% de precisão e 92% de cobertura, de acordo com a avaliação realizada.

Espera-se que o conhecimento gerado na pesquisa, por meio dos *corpora* de CDs e NCDs, das gramáticas locais e da descrição dos padrões verbais definitórios, seja útil

para a Terminologia, PLN e Linguística. Nas três subseções seguintes (11.3, 11.4 e 11.5), trazemos algumas possíveis contribuições desta pesquisa.

11.3 CONTRIBUIÇÕES PARA A TERMINOLOGIA

Dentre as três áreas mencionadas, espera-se principalmente trazer uma pequena contribuição à Terminologia, sobretudo na otimização da etapa da redação da definição terminológica que, como já mencionado, recorre-se a contextos definitório de modo a auxiliar na compreensão do significado do termo que será definido. Com a implementação das gramáticas propostas, é possível ter um sistema que recupere CDs e que possa facilmente ser incorporado a qualquer concordanciador.

Possivelmente as gramáticas locais podem ser usadas também na identificação de relações semânticas com o propósito de auxiliar na construção de ontologias ou mapas conceituais. Outra aplicação possível é utilizar as gramáticas para auxiliar na identificação de unidades terminológicas em *corpus*.

11.4 CONTRIBUIÇÕES PARA O PLN

Os dados produzidos na descrição podem ser proveitosos na geração de mais conhecimento do fenômeno linguístico de forma a serem processados por aprendizado de máquina, em sistemas de pergunta e resposta, sistemas de tradução e na área de Mineração de Textos e Recuperação de Informação em língua portuguesa.

Os *corpora* de CDs e NCDs também podem contribuir com outras investigações que necessitam de descrição linguística para fins computacionais.

11.5 CONTRIBUIÇÕES PARA A LINGUÍSTICA

Toda a tarefa computacional dependente de língua necessita de uma descrição linguística. É claro que, dependendo da tarefa, essa descrição pode ficar nos níveis superficiais (gráfico, morfológico, sintático, léxico) ou nos níveis mais profundos (semântico, pragmático, discursivo, retórico). No caso desta pesquisa, a descrição que fizemos se inclui nos níveis superficiais, já que perpassamos aspectos morfológicos,

sintáticos e léxicos para dar conta da tarefa prevista. Nesse sentido, entendemos que nosso trabalho traz importantes contribuições para a descrição linguística no que concerne aos modos de definir em português. Dito de outro modo, procuramos responder em nossa pesquisa a seguinte pergunta: quando se vai definir em português, como se define e quais estruturas linguísticas são mobilizadas?

Obviamente que existem muito mais realizações linguísticas que revelam um contexto definitório, mas a pequena descrição realizada possibilita compreendermos melhor uma pequena porção da língua, por meio de dados que indicam as estruturas possíveis para elaborar uma definição em texto e a produtividade dessas estruturas nesse gênero textual. A partir da compreensão desse microcosmo da língua, abrem-se várias possibilidades de aplicação desse conhecimento: ensino de escrita de definição em português, como fizeram os linguistas aplicados para o inglês; uso desse conhecimento para ser organizado em dicionários lexicográficos monolíngues e bilíngues e em materiais didáticos.

11.6 TRABALHOS FUTUROS

Como a pesquisa teve como escopo apenas artigos científicos, um trabalho futuro interessante seria a formação de *corpus* a partir de outros gêneros textuais, tais como manuais, artigos jornalísticos, entre outros, a fim de testar os padrões identificados neste trabalho. Outra possibilidade é a implementação da gramática de exclusão e das heurísticas em um classificador de contextos definitórios para a língua portuguesa. E, por fim, sugere-se a análise de outros verbos, como “definir”, “caracterizar”, “formar” e etc.

REFERÊNCIAS BIBLIOGRÁFICAS

ALARCÓN, R. **Descripción y evaluación de un sistema basado en reglas para la extracción automática de contextos definitorios** (Tese de Doutorado). Universidad Pompeu Fabra, Barcelona, 2009.

ALCINA, A.; VALERO DOMÉNECH, E. Análisis de las definiciones del diccionario cerámico científico-práctico. Sugerencias para la elaboración de patrones de definición. In **Debate Terminológico 4**, 2008. Disponível em: <http://seer.ufrgs.br/index.php/riterm/article/view/23841/13830>. Acesso em: fev/2013

ALMEIDA, G. M. B.; KAMIKAWACHI, D.S.; MANFRIN, A.M.P.; SOUZA, I.P.; IZUMIDA, F.H.; FELIPPO, A.D.; ZAUBERAS, R.T.; MELCHIADES, F.G.; BOSCHI, A.O. Glossário de Revestimento Cerâmico. In: Ieda Maria Alves. (Org.). **Cadernos de Terminologia.** 1ed.São Paulo: FFLCH-USP, 2011, v. 4, p. 03-56.

ALMEIDA, G. M. B.; VALE, O.V. Do texto ao termo: interação entre Terminologia, Morfologia e Lingüística de Corpus na extração semi-automática de termos. In: ISQUERDO, A.N.; FINATTO, M.J.B. (Org.). **Ciências do Léxico**. 1ed.Campo Grande: Editora da UFMS, 2010, v. IV, p. 483-499.

ALMEIDA, G.M.B. A Teoria Comunicativa da Terminologia e a sua prática. **Alfa**,

ALMEIDA, G.M.B. O percurso da Terminologia: de atividade prática à consolidação de uma disciplina autônoma. In **TradTerm: Revista do Centro Interdepartamental de Tradução e Terminologia**, v.9, p. 133-134, São Paulo, 2003.

ALMEIDA, G.M.B.; SOUZA, D.S.L.; PINO, D.H.P. A definição nos dicionários especializados: proposta metodológica. In **Debate Terminológico**, v. 3, p. 1-20, 2007. Disponível em: <http://www.riterm.net/revista/n_3/index.htm>. Acesso em: fev/2013.

ALPÍZAR CASTILLO, R. **Cómo hacer un diccionario científico técnico?** Buenos Aires: Editorial Memphis, 1997. 184p.

ALUÍSIO, S.M. **Ferramentas para auxiliar a escrita de artigos científicos em inglês como língua estrangeira**. 1995. 216 f. Tese (Doutorado em Física) – Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos, 1995.
Araraquara, v. 50, p. 81-97, 2006. Disponível em: <http://www.alfa.ibilce.unesp.br/download/v50-2/06-Almeida.pdf>. Acesso em: fev/2013.

ATKINS, P.; JONES, L.; CARACELLI, I. **Princípios de Química**: questionando a vida moderna e o meio ambiente. Bookman, 2001.

ATKINS, S.; RUNDELL, M. **The Oxford guide to practical Lexicograpgy**. Oxford: Oxford University Press, 2008. 540p.

AUGER, A. **Repérage des énonces d'intérêt definitoire dans les bases de données textuelles.** 1997. 224f. Tese (Doutorado em Linguística) - Faculté des Lettres, Université de Neuchâtel, Genebra, 1997. 110p. Disponível em: <http://doc.rero.ch/record/473/files/these_AugerA.pdf> Acesso em: fev/2013

AUGER, P.; ROUSSEAU, L. **Méthodologie de la recherche terminologique**. Québec, Office de Langue Française, 1978.

AULETE. Disponível em: <aulete.uol.com.br/>. Acesso em fev. 2014.

BARROS, L. A. **Curso básico de Terminologia**. São Paulo: EdUSP, 2004. 285p.

BARROS, L. A.; BABINI, M.; AUBERT, F. H. Terminologia e tradução juramentada: questões de tipologia textual e equivalência terminológica interlinguística Português-francês-italiano. In **Filologia e Lingüística Portuguesa**, v. 12, p. 233-249, 2010.

BOULANGER, J.C. Présentation: images et parcours de la socioterminologie. **Meta**,

BOULANGER, J.C. Une lecture socio-culturelle de la terminologie. **Cahiers de linguistique sociale. Rouen**, p. 13-30, 1991.

BOUTIN-QUESNEL, R. et al.**Vocabulaire systématique de la terminologie**. Québec, Publications du Québec – Cahiers de l'Office de la Langue Française, 1985.

BRASCHER, M.; CAFÉ, L. Organização da informação ou organização do conhecimento? In: **Encontro nacional de pesquisa em ciência da informação,** 9. 2008, São Paulo. *Anais*... São Paulo: USP, 2008.1 CD-ROM.Bruxelles: Duculot De Boeck, 2003. 286p.

CABRÉ, M. T. **La terminología: representación y comunicación: elementos para uma teoría de base comunicativa y otros artículos**. Barcelona: IULA/Universitat Pompeu Fabra, 1999. 369p.

CABRÉ, M. T. **La terminología: teoría, metodología, aplicaciones**. Tradução de Carles Tebé. Barcelona: Editorial Antártida/Empúries, 1993. 529p.

CABRÉ, M. T. Theories of terminology: their description, prescription and explanation.
CAO, Y. et al. **Adapting Ranking SVM to Document Retrieval**. SIGIR, 2006.

CASTILLO, R. A. **Cómo hacer un diccionario científico técnico?** Buenos Aires: Editorial Memphis, 1997.

CÉLESTIN, T. et al. **Méthodologie de la recherché terminologique ponctuelle**: Essai de définition. Office de la langue française. 1990. Disponível em : https://www.oqlf.gouv.qc.ca/ressources/bibliotheque/terminologie/recherche_terminolog.pdf. Acesso em 24.07.14.

CIAPUSCIO, G. **Textos especializados y terminología**. Iula, Barcelona, 2003.

COHEN, J. A coefficient of agreement for nominal scales. **Educational and Psychological Measurement,** p. 37-46, 1960.

COLETI, J.S. **Base de dados morfológicos de terminologias do português do Brasil.** Descrição e análise morfológica com vistas à disponibilização on-line. Dissertação (Mestrado em Linguística) - Universidade Federal de São Carlos. 2010.

CONDAMINES, A. Sémantique et corpus spécialisés: constitution de bases de connaissances terminologiques, **Carnets de grammaire**, Relatório interno da ERSS (Équipe de Recherche en Syntaxe et Sémantique), n. 13. Toulouse: CNRS et Université de Toulouse-le Mirail, 2003.

CONFORTO, E. C.; AMARAL, D.C.; SILVA, S.L. Roteiro para Revisão Bibliográfica Sistemática: aplicação no desenvolvimento de produtos e gerenciamento de projetos. In: **8o. Congresso Brasileiro de Gestão de Desenvolvimento de Produto - CBGDP 2011**, Porto Alegre-RS. 8o. Congresso Brasileiro de Gestão de Desenvolvimento de Produto - CBGDP 2011. Porto Alegre: Instituto de Gestão de Desenvolvimento de Produto, 2011.

COUTO, S. **A Definição Terminológica: Problemas teóricos e práticos encontrados na construção de um glossário no domínio da Corrosão**. Porto: FLUP, 2004.

DARIEN, S. The role of definition in scientific and technical writing: forms, functions and properties. In **English Language Research Journal**, p.41-56, 1981.

DE BESSÉ, B. Terminological definitions. Handbook of Terminology Management. Ed. WRIGHTY, S. E., BUDIN, G. Amsterdã, Jonh Benjamins, p.63-74. 1997.

DEL GAUDIO, R.; BRANCO, A. Automatic Extraction of Definitions in Portuguese: A Rule-Based Approach. In **Proceeding of 2nd Workshop on Text Mining and Applications,** Guimarães, Portugal, 2007.

DESMET, I. Questões de semântica em terminologia. A problemática da definição terminológica. **Terminologias**, v. 2, 1990, p. 4-21.

DUBUC, R. **Manual práctico de terminología**. Tradução de Ileana Cabrera. Chile: RiL Editores, 1999. 236p.

DUBUC, R. **Manual pratique de terminologie**. Montréal: Linguatech, 1978.

DUBUC, R.; LAURISTON, A. Terms and Contexts. In **Handbook of Terminology Management.** Volume 1. Wright, S.E. and Budin, G. (eds). Amsterdam/Philadelphia: John Benjamins Publishing Company, 80-87, 1997.

FABER, P. Terminographic definition and concept representation. In **Training the Language Services Provider for the New Millennium**, ed. Belinda Maia, Johann Haller, and Margherita Ulyrich, 343-354. Porto: Universidade do Porto, 2002.

FELBER, H. Manuel de terminologie. Paris, Unesco/InfoTerm, 1984.

FINATTO, M. J. B. **Definição terminológica: fundamentos teórico-metodológicos para sua descrição e explicação**. 2001. 395f. Tese (Doutorado em Letras) - Instituto de Letras, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2001.

FLEIS, J. **Statistical methods for rates and proportions**. New York: John Wiley & Sons, 1981.

FLOWEDER, J. Definitions in Science Lectures**. Applied Linguistics**, p.201-221, 1992.

FONSECA, R.; SILVA, P.; SILVA, R. **Acordo inter-juízes:** O caso do coeficiente kappa. Laboratório de Psicologia, v. 5, n.1, p. 81-90, 207

GAMBIER, Y. Problèmes tertminologiques des pluies acides: pour une socioterminologie. **Meta,** Montréal, v.32, n. 3, p. 314-320, 1987. Disponível em: < http://www.erudit.org/revue/meta/1987/v32/n3/002791ar.pdf>. Acesso em: fev/2013

GAUDIN, F. **Socioterminologie: des problèmes semantiques aux pratiques institutionnelles.** Rouen: Université de Rouen, 1993. 255p. (Publications de l'Université de Rouen, 182).

GAUDIN, F. **Socioterminologie: une approche sociolinguistique de la terminologie**.

GOUADEC, D. **Terminologie: constitution des données**. Paris: AFNOR, 1990. 219p.

HEARST, M. A. TextTiling: segmenting text into multi-paragraph subtopic passages. In **Computational Linguistics**, v.23, n. 1, 1997.

HOFFMAN, L. **Els llenguatges d'especialitat: selecció de textos**. Barcelona: IULA, 1998. 284p. (Sèrie Monografies)

IFTENE, A., TRANDABĂŢ, D., PISTOL, I.: Grammar-based Automatic Extraction of Definitions and Applications for Romanian. In **RANLP 2007 workshop: Natural Language Processing and Knowledge Representation for eLearning Environments**, 2007.

ISO 1087.1. **Terminology work – Vocabulary. Part 1: Theory and application,** 2000.

ISO 12620. **Computer applications in terminology -- Data categories.** 1999.

ISO 704.1 **Terminology work—Principles and methods**, 2000.

KAMIKAWACHI, D. S. L. **Aspectos semânticos da Definição Terminológica (DT): descrição linguística e proposta de Sistematização**. (Dissertação de Mestrado), Departamento de Letras, Universidade Federal de São Carlos, São Carlos, 2009.

KLAVANS, J. L.; MURESAN, S. Evaluation of DEFINDER: A System to Mine Definitions from Consumer-oriented Medical Text. In **Proceedings of JCDL**. 2001.

KLAVANS, J. L.; MURESAN, S. Evaluation of the DEFINDER System for Fully Automatic Glossary Construction. In **Proceedings of AMIA**. 2001.

KOCOUREK, R. **La langue française de la technique et de la science: vers une linguistique d'une langue savante.** Wiesbaden, Brandstette, 1991.

LARIVIÈRE, L. Comment formuler une définition terminologique. **Meta**, Montréal, v. 41, n.3, 405-418, 1996. Disponível em: < http://www.erudit.org/revue/meta/1996/v41/n3/003401ar.pdf> Acesso em: fev/2013

LAVOISIER A. L. **Traité élémentaire de chimie**. Paris, 1789.

MAIA, B.; SARMENTO, L.; SANTOS, D.; CABRAL, L., PINTO, A. CORPÓGRAFO - an online suite of tools for the construction and analysis of corpora, semi-automatic extraction of terminology and the construction of conceptual databases in **Proceedings of Corpus Linguistics Conference Birmingham** 2005, 2005

MALAISÉ, V. **Méthodologie linguistique et terminologique pour la structuration d'ontologies différentielles á partir de corpus textuels.** (Tese de Doutorado). Paris: UFR de Linguistique, Universite Paris 7 – Denis Diderot, 2005.

MATTOS, D.F. **Descrição e análise morfológica da terminologia da Fisioterapia: subsídios para organização de uma base de dados morfológicos de terminologias do português do Brasil**. Dissertação (Mestrado em Linguística) - Universidade Federal de São Carlos, 2013.

MILLER, G. A. WordNet: A Lexical Database for English. **Communications of the ACM.** Vol. 38, No. 11: 39-41, 1995. Montréal, v. 40, n. 2, p. 194-205, 1995. Disponível em: <http://www.erudit.org/revue/meta/1995/v40/n2/002117ar.pdf>. Acesso em fev/2013.

NASCIMENTO, R. F. F et al. **O algoritmo Support Vector Machines (SVM):** Avaliação da separação ótima de classes em imagens CCD-CBERS. Natal, 2009. p. 079-2086.

NAVIGLI, R., VELARDI, P. GlossExtractor: a Web Application to Automatically Create a Domain Glossary. **Proc. of the 10th Congress of the Italian Association for Artificial Intelligence (AI*IA 2007),** Rome, Italy, September 10-13th, 2007, pp. 339-349.

NOY, F. MCGUINNESS D.L. **Ontology Development 101: A Guide to Creating Your First Ontology**. Stanford Knowledge Systems Laboratory Technical Report, 2001.

O Pavel: Curso interativo de terminologia. Disponível em: http://www.bt-tb.tpsgc-pwgsc.gc.ca/btb-pavel.php?page=tdm-toc&lang=fra&contlang=por. Acesso: agosto/2013.

OLIVEIRA JR., C.D. **Extração automática de contextos definitórios em textos acadêmicos da ciência da informação**. Dissertação de Mestrado. Faculdade de Ciência da Informação. Universidade de Brasília. 2012.

PEARSON, J. **Terms in Context**. John Benjamins Publishing Company, 1998.

PEARSON, J. **Terms in context**. John Benjamins Publishing Company. Amsterdam/Philadelphia, 1998.

PINTO A.S.; OLIVEIRA, D. **Extracção de definições no Corpógrafo.** Outubro de 2004. Disponível em: http://comum.rcaap.pt/bitstream/123456789/281/1/OliveiraPintoOut2004.pdf. Acesso em: fev/2013

PRZEPIÓRKOWSKI, A. DEGÓRSKI, L. WÓJTOWICZ, B. On the evaluation of polish definition extraction grammars. In **RANLP 2007 workshop: Natural Language Processing and Knowledge Representation for eLearning Environments**, 2007.

REY, A. **La terminologie: noms et notions**. Paris: Presses Universitaires de France, 1979.

RIZZO, R.P.; LOLLO, J.A. Capacidade de retenção de barreiras de proteção produzidas com solo arenoso estabilizado quimicamente. In **Eng. Sanit. Ambient,** vol.11, p.250-259, 2006. Disponível em: http://www.scielo.br/pdf/esa/v11n3/a08v11n3.pdf. Acesso em: 24.07.14.

RUSSEL, J. **Química Geral,** vol. 2. São Paulo: Makron, 1994.

SAGER, J. C. **Curso práctico sobre el procesamiento de la terminología**. Tradução de Laura C. Moya. Madrid: Fundación Germán Sánchez Ruipérez/Pirámide, 1993. 442p.

SAGGION, H. Identifying Definitions in Text Collections for Question Answering. In: **Proceedings of the 4th International Conference on Language Resources and Evaluation.** Lisboa, p. 1927-1930, 2004.

SAGGION, H; GAIZAUSKAS, R. J. Mining On-line Sources for Definition Knowledge. **FLAIRS Conference** 2004: 61-66

SELINKER, L.; TRIMBLE, R.M.; TRIMBLE, M. On reading english for science and technology: presuppositional rhetorical information in the discourse. In J.C.Richards (ed.), **Teaching English for Science and Technology.** Singapore: Singapore University Press, 37-67, 1976.

SEPPÄLÄ, S. **Composition et formalisation conceptuelles de la definition terminographique.** 2004. 200f. Tese (Doutorado em Tratamento Informático Multilingue), École de traduction et d'interprétation, Université de Genève, Genebra, 2007.

SIERRA MARTÍNEZ, G.; ALARCÓN, R. El rol de las predicaciones verbales en la extracción automática de conceptos. **Estudios de Lingüística Aplicada**, n. 38, 129-144. Distrito Federal, México, 2003.

SIERRA, G.; POZZI, M.; TORRES, J.M. **Proceedings international workshop on definition extraction.** Bulgaria, 2009. Disponível em: http://www.aclweb.org/anthology/W/W09/W09-4400.pdf. Acesso em: 24.07.14.

SUMÉRIA. In: **Wikipédia**, a enciclopédia livre. Flórida: Wikimedia Foundation, 2014. Disponível em: <pt.wikipedia.org/w/index.php?title=Sum%C3%A9ria&oldid=38001530>. Acesso em: 8 fev. 2014.

SWALES, J. **Episodes in ESP**. Oxford, UK: Pergamon Press, 1985.

SWALES, J.M. Definitions in Science and Law: Evidence for subject-specific course components? **Fachsprache**, p. 106-112, 1981.

SWALES, J.M. **Writing Scientific English**. London: Thomas Nelson, 1971.

TEMMERMANN, R. **Towards new ways of terminology description. The sociocognitive approach.** Philadelphia: John Benjamins, 2000. 258p.**Terminology,** v. 9, n. 2, p. 163-200, 2003.

TRIMBLE, L. **English for Science and Technology**: a discourse approach. Cambridge: Cambridge University, 1985.

VÉZINA, R. et al. **La rédaction de définitions terminologiques**. Office québécois de la langue française, Montréal, 2009. Disponível em: https://www.oqlf.gouv.qc.ca/ressources/bibliotheque/terminologie/redaction_def_terminologiques_2009.pdf. Acesso em 24.07.14.

WEBSTER, J.; WATSON, J.T. Analyzing the past to prepare for the future: writing a literature review. **MIS Quarterly & The Society for Information Management,** v.26, n.2, pp.13-23, 2002.

WENDT, Igor da Silveira. **Extração de contextos definitórios a partir de textos em língua portuguesa.** 69f. (dissertação de mestrado). Faculdade de Informática, PUC-RS. Porto Alegre, 2010.

WESTERHOUT, E.; MONACHESI, P. Extraction of Dutch definitory contexts for e-learning purposes. In **CLIN proceedings** 2007, 2007.

WÜSTER, E. **Introducción a la teoría general de la terminología y a la lexicografia terminológica.** Tradução de Anne-Cécile Nokerman. Barcelona: IULA/Universitat Pompeu Fabra, 1998. 227p.

XU, J., CAO, Y., LI, H., ZHAO, M. **Ranking definitions with supervised learning methods**, Special interest tracks and posters of the 14th international conference on World Wide Web, Chiba, Japan, May 10-14, 2005.