Pontifícia Universidade Católica
do Rio de Janeiro

Wograine Evelyn Faria Dias

Dos termos às entidades no domínio de petróleo

Dissertação de Mestrado

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre em Letras/ Estudos da Linguagem pelo Programa de Pós-graduação em Estudos da Linguagem da PUC-Rio.

Orientadora: Prof. Maria Cláudia de Freitas

Rio de Janeiro
Abril de 2021

Pontifícia Universidade Católica
do Rio de Janeiro

Wograine Evelyn Faria Dias

Dos termos às entidades no domínio de petróleo

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre pelo Programa de Pós-graduação em Estudos da Linguagem da PUC-Rio. Aprovada pela Comissão Examinadora abaixo:

Prof. Maria Cláudia de Freitas
Orientadora
Departamento de Letras – PUC-Rio

Prof. Maria Jose Bocorny Finatto
UFRGS

Prof. Diana Maria de Sousa Marques Pinto dos Santos
University Of Oslo

Rio de Janeiro, 26 de abril de 2021

Wograine Evelyn Faria Dias

Graduou-se em letras (Português-Literatura) pela PUC-Rio, em 2018, e concluiu, em 2021, o mestrado, na mesma instituição. Durante o mestrado, foi membro do projeto BIG Oil - Ciência de Dados para Óleo e Gás. O projeto é financiado pela Agência Nacional do Petróleo (ANP) e é realizado em parceria com o Laboratório de Inteligência Computacional Aplicada (ICA), da PUC-Rio. Como pesquisadora, participa de projetos na área de Linguística Computacional. Além disso, é professora de literatura, português e redação.

Ficha Catalográfica

Dias, Wograine Evelyn Faria

Dos termos às entidades no domínio de petróleo / Wograine Evelyn Faria Dias ; orientadora: Maria Cláudia de Freitas. – 2021.
115 f. : il. color. ; 30 cm

Dissertação (mestrado)–Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Letras, 2021.

Inclui bibliografia

1. Letras - Teses. 2. Extração de informação. 3. PLN. 4. Terminologia. 5. Taxonomia. 6. Entidades mencionadas. I. Freitas, Maria Cláudia de. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Letras. III. Título.

CDD:400

Agradecimentos

Ao CNPq e à PUC-Rio, pelos auxílios concedidos, sem os quais este trabalho não poderia ter sido realizado.

À Agência Nacional do Petróleo, Gás Natural e Biocombustíveis (ANP), pelo apoio e incentivo à pesquisa.

Ao Laboratório de Inteligência Computacional Aplicada da PUC-Rio (ICA), que se tornou uma segunda casa ao longo dos dois anos percorridos.

À minha orientadora Cláudia Freitas, pela disponibilidade, pela troca de ideias e pelo convite ao desafio.

Aos PIBICs Elvis, Tatiana e Aline, por todo apoio, crescimento conjunto e parceria.

A todos os pesquisadores do ICA, pelos bons momentos.

A todos os professores, funcionários e alunos do Departamento, pelos ensinamentos e pela ajuda.

Aos professores que participaram da Comissão Examinadora.

Ao meu namorado, Carlos Arthur, pelo incentivo e paciência.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

Resumo

Dias, Wograine Evelyn Faria; Freitas, Maria Cláudia de (Orientadora). **Dos termos às entidades no domínio de petróleo**. Rio de Janeiro, 2021. 115 p. Dissertação de Mestrado - Departamento de Letras, Pontifícia Universidade Católica do Rio de Janeiro.

Este trabalho tem como objetivo identificar uma terminologia e expressões relevantes do domínio de óleo e gás (O&G) e estruturá-la como uma taxonomia, tendo em vista o levantamento de itens para a anotação de entidades dentro do domínio. Para tanto, foi construída uma lista de termos relevantes da área, com base em diversas fontes, e, em seguida, a lista foi estruturada hierarquicamente por meio de regras. O processo de elaboração da taxonomia seguiu aspectos teórico-metodológicos utilizados por diversos trabalhos semelhantes dentro da área. O trabalho procura evidenciar que a identificação de uma terminologia de um domínio técnico e a sua estruturação como taxonomia podem servir como a primeira etapa do levantamento de entidades de um domínio. Por conta disso, o trabalho também se propõe a discutir estratégias para identificação de entidade mencionada (EM) e possibilitar um diálogo entre duas áreas: Processamento de Linguagem Natural (PLN) e Linguística. De maneira geral, espera-se que a taxonomia ajudar a suprir, mesmo que de forma modesta, a escassez de recursos linguísticos para as técnicas do Processamento de Linguagem Natural (PLN) e da Extração de Informação (EI), dentro da área de óleo e gás.

Palavras-chave

Extração de informação; PLN; terminologia; taxonomia; entidades mencionadas.

Abstract

Dias, Wograine Evelyn Faria; Freitas, Maria Cláudia de (Advisior). **From terms to entities in the oil and gas area**. Rio de Janeiro, 2021. 115 p. Dissertação de Mestrado - Departamento de Letras, Pontifícia Universidade Católica do Rio de Janeiro.

This work aims to identify a terminology and relevant expressions of the oil and gas domain and structure it as a taxonomy. To this end, a list of relevant terms in the area was built, based on various sources, and then the list was structured hierarchically by rules. The taxonomy elaboration process followed theoretical and methodological aspects used by several similar works within the area. The work tries to show that the identification of a technical domain terminology and its structuring as a taxonomy can serve as the first stage of the identification of entities in a domain. Because of this, the work also proposes to discuss strategies for identifying named entity and to enable a dialogue between two areas: Natural Language Processing (NLP) and Linguistics. In general, the taxonomy presented is expected to supply, at least in a modest way, the lack of linguistic resources for techniques of Natural Language Processing (NLP) and Information Extraction (EI), within the area of oil and gas.

Keywords

Information extraction; NLP; terminology; taxonomy; named entity.

Sumário

Lista de figuras

Lista de quadros

1
Introdução

A sociedade nunca contou com um volume de informação tão grande como o que há disponível hoje em dia. De fato, vivemos no mundo da transformação digital e nossa capacidade humana de lidar e acompanhar as informações – muitas vezes apresentada pela forma de textos – é parca, diante da quantidade e do movimento de material existente. Por causa disso, diversas técnicas de ciência de dados e inteligência artificial vêm sendo desenvolvidas, com o objetivo de estruturar as informações contidas em textos. Diante disso, investimentos em tecnologias digitais capazes de processar e organizar informação parecem ser cada vez mais necessários.

Do ponto de vista empresarial e industrial, a necessidade de extração de informação pode ser observada em diversos contextos. O trabalho de Cleverley e Burnett (2015) demonstra isso. De acordo com os autores, há desafios, na busca corporativa, para encontrar informações precisas e recuperar informações relevantes, diante do amontoado de informação não estruturada que existe sobre as áreas. Os autores apontam que, segundo executivos, 22% de receita anual é representada por oportunidades perdidas devido a falhas do aproveitamento de informações.

Diversas áreas do conhecimento perdem tempo procurando informação sobre seus setores ou falham em aproveitar as informações de que dispõem. O debate que surge, como consequência dessa deficiência, é justamente em torno da utilização e/ou elaboração de recursos/ferramentas que auxiliem nas tarefas que lidam com organização e processamento de informação.

Diante desse contexto, o objetivo mais específico, e mais prático, deste trabalho é identificar a terminologia de um domínio técnico e estruturá-la como uma taxonomia. Taxonomias são uma maneira de organização do conhecimento, e, por isso, espera-se que uma taxonomia sirva como auxílio em tarefas de recuperação de informação.

Em outras palavras, objetiva-se apresentar uma taxonomia, que sirva como recurso/ferramenta em tarefas de extração de informação. Quanto a isso, esperamos,

por exemplo, que a estrutura sirva como subsídio para o desenvolvimento de um *tagset* (conjunto de etiquetas de anotação) de entidades da área (o que, aliás, já está sendo feito). Com relação ao domínio em que se circunscreve a dissertação, trata-se do domínio do óleo e gás. A principal finalidade deste trabalho é ajudar a suprir, ainda que de maneira modesta, a falta recursos linguísticos para as técnicas do Processamento de Linguagem Natural (PLN) e da Extração de Informação (EI), dentro de um domínio técnico específico. A dissertação se insere no âmbito do projeto Big Oil - Ciências de dados para a indústria de óleo e gás -, realizado com apoio do ICA (Laboratório de Inteligência Computacional Aplicada da PUC-Rio.

Quanto à taxonomia criada, ela foi feita de forma semiautomática, a partir de uma abordagem híbrida e de fontes de diferentes naturezas: glossários e dicionários, por um lado, e um grande *corpus* específico de domínio, por outro. O processo de elaboração do recurso se inspirou em aspectos metodológicos utilizados por diversos trabalhos semelhantes dentro da área e pode ser dividido em duas etapas: na primeira, uma lista de termos relevantes da área foi criada, partindo tanto de terminologias já existentes, como de novos recursos. Na segunda, foi feita uma estruturação da lista criada, com auxílio de regras, com o fim de agrupar termos relacionados e oferecer uma visualização estruturada da lista. Todo este processo será detalhado em páginas seguintes, no entanto, nestas linhas introdutórias, é imprescindível assinalarmos nossa motivação para nos situarmos no domínio técnico, e, especificamente, na área de óleo e gás.

Sobretudo por conta dos avanços da Inteligência Artificial, a comunidade do PLN vem crescendo fortemente com abordagens que pouco ou nada dialogam com os linguistas. No entanto, como aponta Manning (2015), nem todos os problemas que envolvem processamento de língua foram resolvidos somente pelas máquinas e ainda há tarefas especialmente desafiantes para a área, como a Resolução de Correferência, por exemplo, ou os domínios especializados. Uma vez que a presente pesquisa remete a área da Linguística Computacional, e, ao mesmo tempo, se inscreve na Linguística, é imprescindível que que ela possa contribuir em aspectos que ainda se mostram desafiadores, especialmente propondo o apoio de estudos linguísticos, para além do uso de redes neurais.

Sobre a linguagem de domínio, em xeque, nesta pesquisa, não apenas Manning (2015) ilumina a questão – ao destacá-la como um desafio ainda presente na área –, como Cohen et al. (2017) também. Em seu trabalho, os autores destacam

que dentro do domínio biomédico, a correferência, por exemplo, não ocorre da mesma forma que em amostras mais genéricas da língua e que isso afeta os procedimentos técnicos realizados para a anotação. Em suma, não é porque um sistema funciona bem com *corpora* de usos mais gerais da linguagem, que ele dará conta de qualquer tipo de amostra da língua, em qualquer tarefa.

Justamente por isso é que materiais de referência para técnicas do PLN em domínios especializados – *corpora* para avaliação, treino e teste e/ou recursos lexicais ou semânticos, por exemplos – são necessários, pois enquanto as técnicas do PLN podem ser portáveis de domínio para domínio, os materiais de referência, por terem suas singularidades linguísticas, não podem. A necessidade de trabalhos nesse sentido vem continuamente sendo destacada, por exemplo, é válido mencionar o GENIA (Thompson et al., 2017) e o CRAFT (Cohen et al., 2017), *corpora* do domínio biomédico que tiveram suas entidades/termos anotados com base na construção e/ou utilização de ontologias do domínio.

Por conta da falta de *corpora* extensivamente anotados na área biológica é que o GENIA - *corpus* e ontologia de referência na biomedicina - foi criado (Thompson et al., 2017). É dentro deste contexto que a necessidade de as diferentes áreas técnicas desenvolverem recursos lexicais e semânticos de seu domínio pode ser afirmada, pois, de fato, ontologias e outros recursos lexicais podem ser úteis em diversas tarefas, tais como: recuperação de informação, sumarização automática, resolução de anáforas e/ou identificação e classificação de entidades nomeadas.

Agora, de volta ao ambiente corporativo, e do domínio em questão – a área de óleo e gás (O&G) –, o interesse em elaborações de recursos que auxiliem em tarefas que lidam com organização e processamento de informação também pode ser observado. De acordo com Gomes et al. (2018), a indústria vem sendo continuamente desafiada a extrair e aproveitar melhor o conhecimento de suas bases de dados disponíveis.

> Estima-se que uma fração de 80% desses dados sejam armazenados em formatos não estruturados (BLINSTON e BLONDELLE, 2017), pressupondo a existência de valiosas informações possivelmente dispersas em diversas bases não utilizadas em seu completo potencial, ocultas em documentos tais como artigos e estudos técnicos, análises laboratoriais, logs de operação, publicações, periódicos e relatórios diversos. (Gomes et. al, 2018, p. 2)

Justamente porque há demanda dentro da área, é que a presente pesquisa busca oferecer um recurso linguístico que auxilie a organizar o conhecimento desse domínio. Já mencionados anteriormente, Cleverley e Burnett (2015) também se posicionam sobre o setor de óleo e gás. Como já foi apontado, os autores afirmam que organizações de diversos setores desperdiçam grande quantidade de tempo procurando informações e que não conseguem aproveitar bem as informações obtidas. Isto posto, o objetivo do trabalho de ambos é combinar técnicas manuais e automáticas de organização do conhecimento, a fim de melhorar a pesquisa e descoberta de informação dentro do universo do petróleo.

Os autores também justificam a importância desse tipo de pesquisa no contexto específico da indústria de óleo e gás: enquanto os gastos feitos para exploração de petróleo são altíssimos (poços de exploração podem custar mais de 100 milhões de dólares), as chances de sucesso não são tão altas (os poços têm 30% de chance de sucesso). Isso significa que quanto mais informações relevantes e organizadas existirem, mais investimentos inteligentes serão realizados.

No entanto, como indicam pesquisas mencionadas por Cleverley e Burnett (2015), buscar informações leva tempo: 24% do tempo de um profissional de negócios é gasto em busca de informações. No setor de óleo e gás, o número relatado é ainda maior: 40%. Como se não bastasse, 48% das organizações consideram a pesquisa insatisfatória. Uma das razões para as pesquisas corporativas serem consideradas insatisfatórias é o vocabulário utilizado, isto é, a incompatibilidade entre termos de pesquisa e os termos das fontes. De acordo com os autores, duas pessoas não escolhem o mesmo nome para o mesmo conceito 80% das vezes. Justamente por isso é que tesauros e técnicas automatizadas se tornaram uma inspiração para a melhoria desse quadro.

Como resposta a esse cenário, o trabalho faz uma revisão da literatura sobre os diversos métodos de organização do conhecimento – pontuando suas particularidades e similaridades – e faz um estudo de caso, na área do óleo e gás, buscando analisar – por um modelo sinérgico – cada abordagem explicada. Cleverley e Burnett (2015) deixam claro que uma das intenções é desenvolver um modelo que englobe os principais benefícios de cada abordagem em um cenário "o melhor dos dois mundos" (manual e automático).

De acordo com os autores, a contribuição do trabalho seria ajudar a desconstruir o debate 'manual' versus 'automático' em muitas empresas e permitir

que elas desenvolvam estratégias mais eficazes de gerenciamento de informações e conhecimento. Além disso, os autores objetivam aliviar a tensão entre o que geralmente é percebido como abordagens concorrentes e mutuamente exclusivas (CLEVERLEY E BURNETT, 2015).

O trabalho de Cleverley e Burnett (2015) traz algumas considerações: em primeiro lugar, ele denuncia que os antigos distanciamentos entre as áreas interessadas na organização de conhecimentos – Biblioteconomia, Ciência da Informação, Gerenciamento de Dados, RI e Inteligência Artificial – estão se desfazendo e abrindo caminhos de convergência. Ou seja, as áreas podem contribuir, em diálogo, para a estruturação de conhecimentos, em diversos contextos. Além disso, o estudo mostra que é positivo adotar vários métodos (manuais e automatizados) para se extrair informação na área de óleo e gás. Isso pode acelerar o encontro de informações e provocar descobertas. Por fim, no que tange a utilização de taxonomias, os autores colocam:

> A corporate taxonomy language that fits the oil and gas organization is seen as critical to ensuring content governance, navigation and retrieval. This is evidenced by multinationals such as RepsolYPF (SalmadorSanchez and Angeles-Palacios 2008) and Statoil (Munkvold et al. 2006), small independents such as Southwestern Energy (Caballero and Nuernberg 2014) and Apache Energy (Rose 2010), National Oil Companies such as Petronas (Noor and Yassin 2006), service companies such as Baker Hughes (Hubert 2012) and Governments such as the Ministry of Oil and Gas in Oman (Alyahyaee 2012). (CLEVERLEY, P. H. e BURNETT, 2015, sem página)

Diante do contexto apresentado, taxonomias, como vocabulários altamente estruturados, podem ser úteis, como recurso linguístico de referência, no oferecimento de uma terminologia da área – o que ajuda tanto na obtenção dos termos relevantes de um domínio como na anotação de entidades de um *corpus* de linguagem especializada – e ainda como recurso para eliminar ambiguidades e estabelecer relações entre termos, tal como a correferência.

Dado que materiais de referência para serem utilizados no treino de sistemas ou como recurso de tarefas do Processamento de Linguagem Natural (PLN) ainda são por diversas vezes escassos, principalmente no contexto de domínios específicos, e, especialmente, em língua portuguesa, o que justifica e dá relevância a essa pesquisa é que a taxonomia criada, no domínio de óleo e gás, possa: (1) servir como primeira etapa do levantamento de entidades relevantes de domínio – visando

um *tagset* de NER no domínio do óleo e gás –, (2) servir como recurso semântico, material de referência, para técnicas e métodos do PLN, (3) servir como uma forma de organização do conhecimento da área de óleo e gás – e, assim, ser útil no processamento e extração de informação dentro da área – e (4) servir como subsídio para a construção de outros recursos em outros domínios.

De forma concreta, essas aspirações já podem ser observadas, por exemplo, na elaboração de um *tagset* de entidades do domínio de óleo e gás, que está em andamento, a partir da primeira versão da taxonomia apresentada, para anotar as entidades mencionadas em um *corpus* composto por trabalhos acadêmicos sobre petróleo. Com isso, na realidade, esta dissertação também procura evidenciar que a identificação de uma terminologia de um domínio técnico e a sua estruturação como taxonomia podem servir como a primeira etapa do levantamento de entidades de um domínio.

Por conta disso, para além de apresentar a taxonomia, o trabalho possui outros dois objetivos. Primeiro, de maneira mais geral – e por meio de recursos lexicais e semânticos –, o trabalho também se propõe a apresentar e discutir estratégias para identificação de entidade mencionada (EM) em *corpora* de domínios técnicos. Segundo, já de um ponto de vista mais teórico, o objetivo é possibilitar uma interseção e um diálogo entre duas áreas: Processamento de Linguagem Natural (PLN) e Linguística.

Sobre as estratégias de identificar EM, mais adiante, será explicado como o Reconhecimento de Entidades Mencionadas (REM) – uma das tarefas do PLN – é de extrema relevância dentro da área, sendo considerado um primeiro passo da extração de informação. No âmbito desta pesquisa, o que se objetiva é demonstrar como uma taxonomia pode ser útil para a realização desta tarefa, isto é, para a identificação e classificação das entidades. Em paralelo a este tema, quanto à interseção das áreas, ao longo da dissertação, pretende-se mostrar também como estudos linguísticos voltados para terminologia podem estar relacionados – até conceitualmente – com a tarefa de REM (NER, em inglês) dentro de domínios especializados. Ou seja, se, no PLN, pesquisadores procuram as entidades presentes em um conjunto de dados, na Terminologia, os linguistas procuram os termos específicos de uma determinada área.

Ao longo das próximas páginas os três objetivos trazidos nessa introdução serão desenvolvidos. Para conhecimento do leitor, quanto à organização da

dissertação: no capítulo 2, é feita uma revisão bibliográfica das áreas envolvidas na pesquisa, bem como dos temas trazidos. Além disso, os aspectos teórico-metodológicos dos procedimentos envolvidos são destacados. No capítulo 3, apresentamos trabalhos relacionados, que inspiraram o trabalho. No capítulo 4, o processo de elaboração da taxonomia é detalhado em suas duas etapas. O capítulo 5 traz uma análise interpretativa e quantitativa dos resultados obtidos, além de apresentar as melhorias realizadas na taxonomia. Por fim, o capítulo 6 discute todos os pontos trazidos, os pontos a melhorar e como a pesquisa pode ser aproveitada.

2
Fundamentação teórica e metodológica

O ponto de partida de uma dissertação deve ambicionar sobretudo contextualizar o leitor acerca da área e dos tópicos significativos do trabalho. Por isso, antes de mais nada, situamos o presente trabalho dentro das áreas envolvidas, apresentamos conceitos relevantes para a pesquisa – tais como *corpus*, entidade, taxonomia, entre outros – e abordamos aspectos e questões metodológicas. Neste capítulo, em primeiro lugar, explicitamos a noção de terminologia e nosso entendimento de *corpus*. Em seguida, apresentamos a área da Linguística Computacional e questões relacionadas a um posicionamento teórico. Posteriormente, delimitamos a noção de entidade e fazemos uma revisão histórica da tarefa. Depois disso, abordamos a importância dos recursos linguísticos para o PLN, dando especial atenção para "taxonomias". Por fim, versamos sobre questões de ordem metodológica envolvidas na pesquisa.

2.1.
Terminologia

Uma vez que partimos de terminologias para a construção de uma taxonomia de óleo e gás, dentre os recursos linguísticos disponíveis, é válido mencionarmos como as terminologias são relevantes, no universo desta pesquisa. Primeiramente, contudo, é preciso esclarecer com que ideia de terminologia estamos trabalhando, e, para tanto, partimos da obra de Krieger & Finatto (2004).

De acordo com as autoras, a palavra terminologia pode assumir dois grandes sentidos diferentes: ora ela é invocada como um campo de estudos interessado nas unidades lexicais de uma língua de especialidade – e, nesse caso, geralmente vem iniciada por letra maiúscula –, ora ela se refere ao conjunto de termos de uma dada área científica, técnica ou tecnológica. Neste trabalho, a palavra terminologia é tomada, em alguns momentos, pela segunda acepção. No entanto, a ideia de Terminologia, como disciplina, também esbarra nesta pesquisa, uma vez que o

trabalho se situa no diálogo entre áreas distintas: Terminologia e Linguística Computacional (PLN).

Como a área do PLN caminha de forma mais independente da linguística convencional – apoiando-se especialmente na Ciência da Computação e Inteligência Artificial –, a ideia de Terminologia, como campo de estudo, abrange a face mais tradicionalmente linguística do trabalho. Além disso, a ideia de Terminologia – com letra maiúscula – abrange uma área de dupla face: estudo e aplicação (KRIEGER & FINATTO, 2004). Isto é, a Terminologia não é apenas um campo de investigação interessado na expressão de um conhecimento científico, mas também uma área preocupada com a univocidade de uma comunicação especializada no plano interacional. Ou seja, existe um aspecto prático que é o de produzir obras/instrumentos que organizem a informação. Assim, a ideia de Terminologia, como campo de estudo, também aponta para a noção mais aplicada desta pesquisa: organizar um vocabulário especializado e produzir uma taxonomia.

Agora, terminologias, como conjuntos de termos técnico-científicos, estão relacionadas a um tipo de comunicação especializada que possui suas peculiaridades. Por conta disso, elas também podem ser referidas – dentro da Terminologia e área afins (Lexicologia, Lexicografia, Terminografia, etc) – como língua de especialidade (KRIEGER & FINATTO, 2004). Isso significa que terminologias podem, por vezes, expressar um outro tipo de linguagem, diferente da comum, usada em contextos específicos e quase sempre por especialistas de uma área. Nesse trabalho, a linguagem específica em foco seria a do setor de O&G.

Esclarecidas as ideias de "terminologia", cabe agora delimitarmos também o nosso entendimento da palavra "termo". Como explica L'Homme (2020), não existe consenso para a noção de "termo". Entretanto, partindo de Krieger e Finatto (2004), vamos assumir que termo é "simultaneamente, elemento constitutivo da produção do saber, quanto componente linguístico, cujas propriedades favorecem a univocidade da comunicação especializada" (KRIEGER & FINATTO, 2004, p. 75). Pelas mesmas vias, também podemos dizer que o termo é o objeto primordial da grande área Terminologia, que, por sua vez, conta com mais outros dois objetos: a fraseologia (expressões idiomáticas) e a definição (enunciados que explicam os termos).

> O léxico especializado apresenta um caráter multidimensional e poliédrico, podendo ser concebido a partir de três dimensões distintas: da linguística; da filosofia; e das diferentes disciplinas científicas. Do ponto de vista linguístico, o termo é visto como uma unidade de significação; para a filosofia é uma unidade de conhecimento; e, para as diferentes áreas do conhecimento é uma unidade de representação. Ou seja, o termo é uma unidade de conhecimento de um domínio de especialidade; é uma unidade de comunicação e divulgação do conhecimento científico; e, é uma unidade lexical. (VAN DER LAAN, 2002, p. 48)

Ao afirmarmos que termo pode ser tanto unidade linguística, como de conhecimento, esbaramos em uma discussão relevante dos estudos de terminologia. Tradicionalmente, o termo foi visto como uma unidade de conhecimento, e não como uma unidade linguística. Isto significa que na área não havia espaço de confusão entre palavra ortográfica e termo, tampouco os termos podiam ser atingidos por polissemia. De forma convencional, os nomes dos termos eram vistos "como meros rótulos e etiquetas com as quais, conscientemente, denominam-se os resultados das ciências e das técnicas (KRIEGER & FINATTO, 2004, p. 78). Ao longo dos anos, a Terminologia também foi influenciada por abordagens mais pragmáticas e interessadas no uso da língua. A partir de uma perspectiva menos tradicional, o termo passou a ser visto como uma entidade multifacetada e complexa.

De todo modo, para além de sua natureza, a crença de que o termo, quando usado recorrentemente, garante a univocidade da comunicação especializada se manteve sólida. O termo, assim, realiza duas funções essenciais: a de representar e a de transmitir o conhecimento especializado (KRIEGER & FINATTO, 2004).

Todavia, entender o que delimita ou confere estatuto de termo a uma palavra é tão ou mais importante que a sua definição teórica. No entanto, apresentar estas condições não é algo tão simples, porque, como foi mencionado, há uma diversidade de posicionamentos sobre a natureza do termo. De acordo com Krieger e Finatto (2004), ainda com este cenário, seria possível elencar traços distintivos dos termos (que os diferem de palavras comuns), esses traços são: (1) tendência a composição sintagmática; (2) tendência a pertencer a categorias gramaticais nominais e (3) arbitrariedade tênue.

Sobre o primeiro, os termos podem ser simples e compostos, mas há uma predominância maior de unidades complexas (sintagmas compostos) nas terminologias. Quanto ao segundo, os termos geralmente são nomes, isto é, pertencem a classe de substantivos, e menos frequentemente podem pertencer a

outras classes, como adjetivos ou verbos. Com relação ao terceiro, há evidências de que há componentes morfológicos – radicais, afixos e sufixos – altamente produtivos para a formação do léxico especializado, o que demonstra que o termo pode não ser tão arbitrário. Além dessas características, o termo também apresenta configurações sígnicas, podendo abranger siglas, acrônimos, abreviaturas e fórmulas.

Por fim, ainda sobre Terminologia, é válido mencionarmos que mais recentemente o uso de *corpora* e de ferramentas tecnológicas que automatizam o processo de produção de terminologias tem se tornado comum na área, e, inclusive, já existe uma especialização voltada para isso: a Terminologia Computacional. No que concerne as vantagens no uso dessas tecnologias, Krieger & Finatto (2004) apontam:

> De fato, a contribuição da Terminologia Computacional tem tornado viável a investigação das linguagens especializadas em uma grande extensão de documentos, além de agilizar a coleta e seleção de termos em grandes volumes de texto. Nessa direção, temos também a criação de bases de dados que podem ser continuamente gerenciadas e atualizadas, com as quais se facilita a editoração de dicionários e de outros produtos impressos, visto que as bases que os geram costumam estar acopladas a bases de textos. (KRIEGER & FINATTO, 2004, p. 140)

Já como desafios no uso da tecnologia, temos o fato de que as ferramentas possuem também a sua margem de erro e podem apresentar desempenhos melhores ou piores. Na tarefa de extração automática de termos, em corpus de domínios especializados, por exemplo, ferramentas de anotação morfossintática podem chegar a uma margem de índice de erro de até 20%: “esse percentual de erro é natural, afinal a marcação, morfossintática envolve vários fatores e variáveis que precisam ser simultaneamente reconhecidos por um programa” (KRIEGER & FINATTO, 2004).

Essa realidade é relevante, já que, neste trabalho, para listarmos candidatos a termos relevantes e/ou relacionados ao domínio de óleo e gás, partimos tanto de nomenclaturas e terminologias prontas (glossários e dicionários), como do reconhecimento terminológico automático. Vamos chegar a fazer a descrição de todo esse processo, no entanto, antes, é válida, uma contextualização acerca das noções de *corpus* e PLN.

2.2.
Corpus e Linguística Computacional / PLN

A noção de *corpus* é bastante importante e comum no PLN, mas é possível dizer que ela é ainda mais importante aos linguistas, porque o *corpus* funciona como mediador entre as tarefas computacionais e o trabalho do linguista. Enquanto cientistas da computação ou programadores, inseridos no PLN, podem, por vezes, trabalhar com o que eles chamam de "*dataset*", o linguista – no PLN – normalmente trabalha com o *corpus*, um conjunto de dados linguísticos. Tendo em vista esta realidade, e o fato de este trabalho partir da Linguística, é fundamental delimitarmos a noção de *corpus*, entendida no escopo da pesquisa.

Dentro da área, podemos encontrar visões diferentes e conflituosas para o conceito de "*corpus*" (SANTOS, 2008; Mc ENERY e HARDIE, 2012). Por vezes, podemos encontrar a ideia de *corpus* como uma área de estudo própria e independente, ainda que focada em métodos e procedimentos para estudar a língua. Nessa via, *corpus* também pode ser entendido como sendo o "objeto" de uma área, da "Linguística de *Corpus*".

Uma outra forma de se entender *corpus* é como uma ferramenta com a qual fazemos linguística. Dentro desta visão, afirma Santos: "Um corpo é uma colecção classificada de objectos linguísticos para uso em Processamento de Linguagem Natural/Linguística Computacional/Linguística" (SANTOS, 2008, p. 45).

É com esta última visão que este trabalho se alinha. Ou seja, no escopo desta pesquisa, entendemos *corpus* como um conjunto de textos que pode ser classificado por pessoas e por máquina e que nos permite estudar a língua. Tal conjunto pode ainda contar com um dispositivo de acesso, anotação e consulta. De tal modo, *corpus* não seria uma área ou o objeto de uma área, mas uma ferramenta de investigação: "O meu ponto de partida é o de que um corpo não é o objecto de estudo do que em inglês se chama *corpus linguistics*, mas sim a ferramenta, o utensílio com que se faz linguística, por isso a minha denominação "linguística com corpos"[*] (SANTOS, 2008, p. 46).

O *corpus*, a anotação e a ferramenta de busca (consulta) são, para os linguistas, o que torna possível explorar questões de pesquisa que seriam

[*] Aqui a autora está usando a expressão "linguística com corpos" em substituição à Linguística de *Corpus*, para marcar que não se trata de uma área independente, com seu objeto de estudo, mas de uma maneira de investigar a língua.

inimagináveis de outra forma dentro da Linguística (Mc ENERY e HARDIE, 2012). O *corpus* oferece uma amostra da língua que – por ser muito grande – não poderia ser estudada a olho nu. Já a anotação permite ao linguista pensar em perguntas para fazer a esse *corpus*. Por exemplo, a anotação gramatical e sintática permite a pesquisa de um fenômeno específico na língua. Quanto às ferramentas, elas são poderosos auxílios para os linguistas, pois são justamente o meio pelo qual o linguista fará as suas perguntas. Entretanto, elas também limitam e definem o que se pode fazer em um *corpus*.

Circunscrita nossa definição de *corpus*, cabe agora, mais do que nunca, mencionarmos também o que é possível fazer com ele. Se o *corpus* é uma coletânea de textos e uma ferramenta, o que podemos fazer com ela? De acordo com Santos (2008), *corpora* possuem quatro tipos de uso, eles podem ser usados: (1) para se ter uma ideia de um determinado problema ou aspecto da língua; (2) para medir um determinado fenômeno; (3) para avaliar uma determinada hipótese ou teoria e (4) para criar outras coisas, tais como: estruturas de conhecimento (dicionários, terminologias, ontologias, etc), materiais didáticos, sistemas de detecção de plágio, etc. No âmbito desta pesquisa, a evidente utilidade de um *corpus* de óleo e gás tem relação com o uso 4, já que, efetivamente, para a criação da taxonomia apresentada, partimos de um *corpus* de textos da área.

Agora, se situamos esta pesquisa na Terminologia, também a situamos, não somente em estudos linguísticos que fazem uso de *corpus*, mas também na Linguística Computacional – ou PLN (Processamento de Linguagem Natural). Em poucas palavras, a Linguística Computacional é um campo de investigação interessado no processamento computacional da linguagem humana (MITKOV, 2003), isto é, é uma área de estudo interessada nos contextos em que máquinas precisam lidar com – para não dizer "entender" – a linguagem humana.

Algumas das tarefas mais comuns no campo da Linguística Computacional são: tokenização do texto; anotação de part-of-speech (POS) – anotação gramatical –; parsing – anotação sintática –; identificação e classificação de entidades mencionadas (REM); resolução de anáfora e/ou correferência; sumarização e anotação de expressão multi-palavra (MWE) (MITKOV, 2003). As tarefas, em grande maioria, estão relacionadas com anotação linguística de grandes bases de dados, que, em nosso domínio, chamamos de *corpus*. Tanto por utilizar corpus,

como por discutir a noção de entidade mencionada, é que a inserção nessas áreas pode ser ressaltada, quando situamos a presente pesquisa.

Atualmente, do ponto de vista metodológico, as pesquisas em PLN geralmente funcionam a partir de duas abordagens diferentes: aprendizado de máquina (*Machine Learning*) e regras linguísticas. No entanto, elas não são necessariamente excludentes e os modelos podem ser híbridos. A função dos programadores ou dos cientistas da computação é criar sistemas baseados nessas abordagens para dar conta de questões e tarefas do PLN.

O aprendizado de máquina (ML) – referido como uma subárea da Inteligência Artificial – requer uma grande quantidade de dados para ser eficaz. Nessa abordagem, os algoritmos "aprendem" pela experiência, isto é, eles extraem padrões a partir de um conjunto de dados e são capazes de generalizar a partir de exemplos (ALLENDE-CID, 2019). Desse modo, o desempenho dos algoritmos melhora à medida que são expostos a mais dados.

Dentro dessa abordagem, existe o aprendizado profundo (*Deep Learning),* uma subárea do ML, que ocupa o lugar central das inovações em inteligência artificial. Essa técnica consiste no treinamento de redes neurais artificiais e ela vem permitindo níveis de aprendizado (abstrações), do ponto de vista da máquina, cada vez mais altos (BASSANI, 2019). Por exemplo, o reconhecimento de faces, o reconhecimento da fala e a identificação de imagens são tarefas típicas desse método. Por sua vez, ele, que é o mais em voga atualmente, pouco ou nada dialoga com a Linguística.

> Rede Neural Computacional é uma técnica que mostrou grande potencial no campo de Machine Learning. A técnica é feita aplicando uma série de camadas que atuam de maneira análoga a um neurônio, executando o processamento de uma pequena parte da informação total. Deep Learning é a aplicação de uma quantidade massiva de camadas de processamento em um algoritmo de rede neural (COPELAND, 2016). (...)
> Deep Learning é o estilo de aprendizagem de máquina que se faz com rede neural profunda, em essência, uma percepção apurada de inteligência artificial, que se parece com a do ser humano e é capaz de gerar conteúdos baseada no aprendizado a partir dessa assimilação. Os algoritmos de DL são capazes de analisar dados não-estruturados sem que haja algum tipo de pré-processamento ou supervisão (GOODFELLOW, BENGIO, COURVILLE, 2016). (PACHECO & PEREIRA, 2018, p. 34-35)

Já a abordagem de regras, consiste na aplicação de uma série de filtros/regras linguísticas. Nesse método, quando as condições das regras são

satisfeitas, isto é, quando a regra se apresenta verdadeira, é possível extrair informações a partir de um banco de dados ou fazer anotações de forma automática (MITKOV, 2003). Abordagens baseadas em regras requerem conhecimento linguístico prévio com relação à língua a ser tradada.

Para entendermos ainda melhor sobre a área da Linguística Computacional, é válido fazermos, muito brevemente, uma revisão histórica do campo. Karen Spärck Jones (1994), em poucas palavras, divide a história da Linguística Computacional em quatro fases:

> "The first phase of work in NLP as lasting from the late1940s to the late 1960s, the second from the late 60s to the late 70s and the third to the late 80s,with the fourth phase to the end of the century. The first phase was driven by MT (machine translation), the second flavoured by AI (artificial inteligence), the third grammatico-logical,while the fourth phase has focused on lexical and *corpus* data". (SPÄRCK JONES, 1994 /2, p. 2)

Como podemos observar, a Linguística Computacional é uma área que surgiu em meados do século XX, inicialmente, com interesse em tradução automática. Mais para frente, ela foi influenciada pelo surgimento da inteligência artificial e, por conta da complexidade da linguagem em uso, o conhecimento de mundo e seu papel na construção e manipulação de representações de significado ganharam bastante atenção.

Como explica a autora, se, na segunda fase, o PLN foi orientado pela semântica, na terceira, durante a década de 80, foi orientado pela gramática. Essa fase pode ser descrita como gramatical-lógica e foi estimulada tanto por estudos linguísticos, que, na época, desenvolveram teorias gramaticais, quanto pelo crescimento da programação lógica. Por fim, o PLN se voltou para uma abordagem mais lexical e para o uso de *corpora*. Na quarta fase, houve uma tendência para abordagens probabilísticas e interesse em sistemas multimodais.

A revisão histórica de Spärck Jones (1994) começa a partir de 1940 e vai até o fim do século XX. Isso significa que a autora não reflete sobre acontecimentos mais recentes da Linguística Computacional. No entanto, apesar de o trabalho de Spärck Jones (1994) não apresentar um cenário ainda mais atual do PLN, a autora reflete sobre a relação, por vezes próxima, por vezes distante, entre a Linguística Computacional e os linguistas. E quanto a isso, vale ainda escrever algumas linhas.

No início, o PLN começou com a colaboração dos linguistas e ambas as áreas, de algum jeito, caminhavam juntas, compartilhando algumas crenças sobre a linguagem. No entanto, os modos de se investigar a língua, com o passar dos anos foi se opondo. Ao longo do século XX, enquanto, na Linguística, houve um intenso investimento em pesquisas baseadas em ideais chomskianos, na Linguística Computacional, o estudo da língua como possuindo uma estrutura universal, ou sistema generalizável, ou objeto idealizável deixou de fazer sentido, pois o PLN precisa dar conta de questões práticas que envolvem a linguagem humana, isto é, ele está voltado para a linguagem em uso, e não para uma língua ou falante ideal (FREITAS, 2017).

Os legados de Saussure e Chomsky para a Linguística são de importância já reconhecida. De fato, essa dupla dominou os estudos linguísticos do século passado e até os dias de hoje ambos possuem relevância teórica expressiva dentro da área. No entanto, se podemos dizer que Chomsky contribuiu fortemente para a ascensão meteórica da Linguística como disciplina, depois da Segunda Guerra Mundial, também podemos afirmar que ele colaborou para que o PLN seguisse de forma independente e sem se referir à Linguística mais convencional (MANNING e SCHÜTZE, 1999).

> (...) The crucial problem is that computational linguistics is fundamentally, and essentially, about process, about working with language, and linguistics as she is spoke is not about language processing. Just as in computation in general, process is not the mere implementation of some non-process account of something, but supplies the motivational account. Thus in computational linguistics, the mechanisms for selecting word senses provide interpretations for what it is for a word to have a sense in relation to a linguistic structure, and for what a linguistic structure is when built by disambiguation. Linguistics in its mainstream forms is not about process, namely, about algorithmic process: Computational linguistics is about process, and in a more thorough sense not just than Chomskyan performance, but than "computation" as a generic abstraction. (SPÄRCK JONES, 2007, p. 438)

É válido mencionar que Chomsky, ao longo do século XX, também se posicionou contra o estudo da língua por meio de *corpus*. O linguista fez críticas a essa abordagem, pondo em dúvida seu caráter científico e questionando a possibilidade de se obter bons resultados (Mc ENERY e HARDIE, 2012). Na metade do século XX, o impacto de sua pesquisa, de postura racionalista, foi tão grande que o estudo da língua por observação foi posto de lado. Com o passar do tempo, o uso de *corpora* não só se mostrou relevante para o PLN, como também

para diversas áreas da Linguística, tais como: aprendizado de línguas estrangeiras, análise do discurso, teorias linguísticas, descrição da língua, lexicografia, terminologia, gramática etc.

Como coloca Spärck Jones (2007), a partir da década de 70, o PLN começou a caminhar por outra direção, deixando de dialogar com a Linguística hegemônica e se aproximando da utilização de *corpus.* No final da década, a Linguística Computacional já parecia avançar de forma independente e a relação com os linguistas minguou ainda mais com o advento do *Machine Learning* e, em seguida, do *Deep Learning*.

Entretanto, autores como Manning (2015) e Spärck Jones (2007) estão longe de afirmarem que a comunidade PLN não pode ou não deve se beneficiar dos linguistas, apesar de que a área já mostrou não precisar se referir à linguística mais convencional. Na verdade, ambos fazem o oposto: esperam que linguistas possam contribuir em pontos que ainda se mostram desafiantes para o campo. É nesse aspecto que a dissertação objetiva contribuir. Além disso, em via de mão dupla, a Linguística também pode se beneficiar do PLN, para estudar a língua a partir de outras abordagens.

2.3. Entidades Mencionadas no PLN

O objetivo mais geral desta pesquisa é estruturar um vocabulário de óleo e gás de modo a apresentar uma taxonomia. Por outro lado, uma das motivações do trabalho é oferecer à comunidade do PLN recursos linguísticos que auxiliem em tarefas da área. É justamente por isso que também debatemos nesta dissertação sobre estratégias para se obter entidades mencionadas em *corpora* de domínio técnicos, pois o processo de construção da taxonomia não só trouxe reflexões sobre o que seriam as entidades em domínios técnicos, como a própria taxonomia em si se tornou um bom ponto de partida para se realizar essa anotação. Nesta seção, apresentamos melhor o conceito de entidade, bem como sua trajetória e relevância no PLN.

A noção de "Entidade Mencionada" (EM), está estritamente relacionada à tarefa de identificação e classificação de entidades[*] (REM), primordial dentro da Linguística Computacional. Esse termo, de fato, surgiu no âmbito de aplicações de Processamento de Linguagem Natural (PLN) e em tarefas de Extração de Informação (EI). Foi no MUC – uma avaliação conjunta de sistemas da área – que se propôs, pela primeira vez, em sua sexta edição, que uma tarefa de reconhecimento de entidades no *corpus* fosse medida de forma independente, pois, durante anos, o reconhecimento de entidades foi entendido como tarefa mais básica de extração de informação de textos (SANTOS & CARDOSO, 2008).

Vislumbrar a pertinência das entidades mencionadas dentro do PLN, nos leva a pensar na definição deste conceito. Entretanto, a noção de "Entidade Mencionada" (EM), como já foi dito, possui relação direta com a tarefa de reconhecimento de entidades, o que nos leva a crer que pensar em uma significação para o termo é entender o que a tarefa de identificação de entidades mencionadas engloba.

Quanto a isso, temos que o objetivo das tarefas de identificação e classificação das entidades é o de identificar e distribuir – a partir de um texto – palavras que representam entidades em categorias predefinidas. Por exemplo, na frase "*U.N official Ekeus heads for Baghdad*" (O oficial da ONU Ekeus segue para Bagdá), temos que as palavras "U.N", "Ekeus" e "Baghdad" podem ser identificadas como entidades e classificadas como EM do tipo organização, pessoa e lugar, respectivamente (TJON KIM SANG e FIEN DE MEULDER, 2003).

De outro modo, a tarefa de REM (Named Entity Recognition - NER) serve basicamente para estruturar a informação presente em grandes volumes de textos. A sua extrema relevância para o PLN se deve a isso: por estruturarem informação, as entidades são como o primeiro passo substancial da extração e processamento de informação. De mais a mais, a tarefa de REM serve de suporte para diversas outras técnicas e aplicações, tais como: *question answering*, *information retrieval*, *co-reference resolution*, *topic modeling* etc (YADAV e BETHARD, 2018).

Como colocam Santos & Cardoso (2007), REM integra a maioria dos sistemas inteligentes que processam língua e a qualidade de sua anotação interfere nos resultados da extração de informação e na realização de outras tarefas. Não é

[*] E identificar e classificar podem tanto ser entendidas como tarefas separadas voltadas para entidades, como partes da mesma tarefa.

exagero afirmar que diversas outras tarefas do PLN dependem da anotação de entidade.

Como já foi explicitado, as entidades podem ser entendidas como unidades linguísticas portadoras de informação que oferecem pistas importantes sobre o que está sendo dito em um *corpus*/domínio. Elas são fundamentais – quase como premissas – se o que se deseja é extrair informação: pois, na medida em que diferenciam "pessoas" de "organizações" ou "lugares", por exemplo, organizam – dão nome – às coisas. Por falar em nome, cabe destacar aqui o fato de que, inicialmente, a tarefa de reconhecimento das entidades mencionadas consistia basicamente na identificação de nomes próprios (SANTOS & CARDOSO, 2007).

Até hoje, podemos entender entidades como unidades de informação, em geral, de caráter nominal. No entanto, o que vale como EM pode ser dependente de projetos, *corpora* e domínios. Se, no início, o trabalho com REM era feito com poucas categorias e de modo mais restrito, com o passar do tempo, as categorias de REM se expandiram significativamente, gerando novos desafios para a área (SANTOS & CARDOSO, 2007). Considerando isso, talvez uma forma mais interessante de se compreender com mais aprofundamento o termo "entidade mencionada" seja através de uma breve, e injusta – é claro –, revisão da literatura disponível, o que significa, neste caso, olhar para algumas das competições que avaliaram sistemas capazes de identificar e classificar entidades mencionadas.

2.3.1. As competições de sistemas de REM

Nesta seção, é feita uma revisão das principais competições voltadas para REM dentro do PLN. Ao longo do texto, buscamos especialmente dissertar sobre dois tópicos: (1) o que são as entidades – isto é, o que conta como entidade em cada competição – e (2) quais são as classes/tipos de entidade envolvidas. De mais a mais, outras questões foram levantadas, com a finalidade de contextualizar o leitor.

A primeira vez que o termo "Entidade Mencionada" apareceu foi no MUC (Message Understanding Conference), precisamente, em sua sexta conferência (MUC-6), em 1995. Financiada pela DARPA[*], o objetivo do MUC era desenvolver

[*] Defense Advanced Research Projects Agency é uma agência de projetos de pesquisa criada com o fim de manter a superioridade tecnológica dos EUA, contra adversários estrangeiros potenciais, como a União Soviética.

e aprimorar métodos para extrair informação. Apesar de se tratar de uma conferência – como o nome já acusa –, a característica mais marcante do MUC era a competição que ele, em cada edição, estimulava (GRISHMAN e SUNDHEIM, 1996).

Resumidamente, os participantes eram motivados a inscreverem seus sistemas nessas competições, caso quisessem participar da conferência. Assim, em cada edição, os grupos inscritos recebiam instruções sobre o tipo de informação a ser extraída e os participantes tinham que desenvolver um sistema que desse conta de atingir o objetivo da tarefa. Até então, ocorreram sete edições da conferência, entre 1987 e 1997. Cada edição trabalhou dentro de um domínio específico e sugeriu uma tarefa diferente para os sistemas, a tarefa de identificação de entidades mencionadas foi adicionada na sexta edição do Message Understanding Conference. Nesta edição, os participantes utilizaram textos do Wall Street Journal como material de teste para a realização da tarefa.

> The first goal was to identity, from the component technologies being developed for information extraction, functions which would be of practical use, would be largely domain independent, and could in the near term be performed automatically with high accuracy. To meet this goal the committee developed the "named entity" task, which basically involves identifying the names of all the people, organizations, and geographic locations in a text. (GRISHMAN e SUNDHEIM, 1996, p. 467)

Inicialmente, como já foi mencionado, as entidades mencionadas consideradas marcáveis representavam, de modo geral, os nomes próprios dos textos. Como se percebe, pelo fragmento acima, as categorias de entidade utilizadas no MUC foram (1) nome de pessoas, (2) organizações e (3) lugares. Essa foi a primeira forma das entidades mencionadas. Em adição a isso, também podiam ser marcadas expressões temporais, como datas e horas e expressões numéricas, monetárias e/ou percentuais. Qualquer coisa, para além disso, a princípio, não contava.

Durante a tarefa, verificou-se um problema relacionado às entidades que, ao longo das competições, ganhou diversos tipos de direcionamentos: é a questão da segmentação e delimitação da entidade. Isto tem a ver com quando uma determinada entidade pode ser considerada como encaixada em outra. Por exemplo, na entidade “Universidade Federal do Rio de Janeiro”, “Rio de Janeiro” tanto pode ser lido como outra entidade do tipo lugar, como parte da entidade “Universidade”,

um de tipo organização. Esse tipo de desafio ficou conhecido na área como entidades embutidas. No MUC, quando elas ocorreram, ambas as possibilidades de respostas foram consideradas como corretas (SANTOS & CARDOSO 2007).

A tarefa de REM excedeu as expectativas e foi um sucesso em termos de resultado. A maior parte dos sistemas obteve uma avaliação acima de 90%, sendo que o melhor sistema atingiu 96% de abrangência e 97% de precisão. O desempenho dos sistemas foi avaliado considerando (1) se a classificação estava correta, ou seja, se o tipo de entidade que se previa estava correto e se (2) os limites da entidade – a delimitação – também correspondia ao esperado. Sobre a delimitação das entidades, vale destacar que o MUC não trabalhou com base na combinação exata, isto é, a inclusão ou não de artigos e preposições não afetava o resultado, contanto que o substantivo próprio estivesse correto (JIANG et al., 2016).

A precisão foi definida como o número de entidades que um sistema previu corretamente, dividido pelo número de todas as entidades que o sistema previu. Já a abrangência foi definida como o número de entidades que um sistema previu corretamente dividido pelo número de entidades identificadas pelos anotadores humanos. Por fim, a micro f-score – nota final – foi definida como a média harmônica de precisão e abrangência (YADAV e BETHARD, 2018).

Sem dúvida, o MUC representou a largada inicial no desenvolvimento de competições que avaliam o reconhecimento e a classificação de entidades mencionadas, pois foi a primeira vez que houve um ambiente de comparação entre sistemas. A partir do MUC, surgiram várias outras avaliações conjuntas, por exemplos: o MET (Multilingual Entity Tracking) – uma adaptação da MUC para japonês, espanhol e chinês – e a CoNLL (Conference on Computational Natural Language Learning), que começou em 1999 e permanece desenvolvendo avaliações até hoje.

Assim como o MUC, o CoNLL é uma conferência marcada por competições anuais. Em cada edição, uma tarefa específica – dentro da área de PLN –, é requerida aos sistemas que desejam participar da conferência. No caso do CoNLL, as edições que desenvolveram tarefas ligadas à detecção de entidade mencionada ocorreram em 2002 e 2003.

As duas, CoNLL-2002 e CoNLL-2003, convergem em alguns pontos: primeiro, ambas trabalharam apenas com quatro tipos de entidade: pessoa, local, organização e miscelânea, sendo que esta última foi criada para englobar tudo o que

não se classificasse nas outras três. Segundo, em ambas, os sistemas participantes foram encorajados a incluir algum componente de aprendizado de máquina (*Machine Learning*). E, por fim, nas duas havia o interesse por parte dos organizadores de que os participantes utilizassem abordagens que fizessem uso de outros recursos, para além do material de treino. Ou seja, a utilização de materiais com informação não anotada, tais como nomenclaturas, dicionários geográficos ou índice de topónimos podiam ser usados para ajudar a resolver a tarefa.

O CoNLL-2002 trabalhou com duas línguas: espanhol e holandês. O material de treino e teste, tanto da língua espanhola quanto da holandesa, foi constituído por artigos jornalísticos. O material da língua holandesa, particularmente, possuía anotação de classe gramatical. Doze sistemas participaram e o que obteve melhor resultado em espanhol atingiu uma score de 81,39%, já no holandês, o sistema que se saiu alcançou 77,05% (TJONG KIM SANG, 2002).

O CoNLL-2003 trabalhou com inglês e alemão. O material de treino e teste de inglês era constituído por textos de um *corpus* de uma agência de notícias britânica, enquanto o de alemão eram textos jornalísticos alemães retirados de um *corpus* multilíngue. No CoNLL-2003 dezesseis sistemas participaram. Em inglês, o sistema que se saiu melhor atingiu 88,76% e, em alemão, o melhor sistema atingiu 72,41% (TJONG KIM SANG & DE MEULDER, 2003).

Quanto a forma de avaliação, no CoNLL, a identificação de entidade mencionada só foi considerada correta quando a marcação do sistema se equivaleu totalmente com a do gabarito, isto é, como o previsto. Isto significa que, em termos de segmentação, se o sistema marcasse “Estados Unidos”, mas o gabarito fosse “os Estados Unidos”, a resposta estaria incorreta. Além disso, no caso de entidades mencionadas embutidas, isto é, quando uma entidade se situava dentro de outra maior, apenas a maior delas deveria/poderia ser marcada (YADAV e BETHARD, 2018).

Também em 1999, teve início outro projeto voltado para reconhecimento de entidade mencionada: o ACE (Automatic Content Extraction). O programa tinha como objetivo desenvolver tecnologias para extrair automaticamente informações da linguagem humana. Até 2001, os esforços se voltaram apenas para tarefas de detecção de entidade. De 2002 a 2003, foi explorada a tarefa de identificação de relações entre entidades. Já em 2004, a tarefa de reconhecimento de eventos foi adicionada. É sobre esta última fase que a seção foca, principalmente porque o

programa ACE, amadurecido em 2004, propõe um olhar mais amplo sobre as entidades mencionadas (DODDINGTON et al., 2004).

O ACE-2004 abrangeu três línguas, inglês, chinês e árabe. Além disso, para a edição, três tarefas principais foram definidas: (1) reconhecimento de entidades, (2) reconhecimento de relações entre entidades e (3) extração de eventos. Os anotadores do programa, que tinham sólido conhecimento linguístico, produziram material de treino e de avaliação, que, nesse caso, foi constituído por textos jornalísticos. Os anotadores, além de etiquetarem as três tarefas, também receberam uma quarta tarefa de anotação: "entity linking" – sobre correferência entre entidades.

A grande contribuição do ACE, quanto a REM, está no seu entendimento do que conta como entidade mencionada. Nesse programa, como entidades, não foram consideradas apenas os nomes próprios dos textos, mas tudo que equivalesse a um determinado ser – seja esse ser humano e animado ou não. Para o ACE, a entidade não se apresenta somente na forma de nome próprio, portanto, todas as maneiras pelas quais uma entidade pode ser referida deve ser considerada.

> (…) In general objective, the ACE program is motivated by and addresses the same issues as the MUC program that preceded it (NIST 1999). The ACE program, however, attempts to take the task "off the page" in the sense that the research objectives are defined in terms of the target objects (i.e., the entities, the relations, and the events) rather than in terms of the words in the text. For example, the so-called "named entity" task, as defined in MUC, is to identify those words (on the page) that are names of entities. In ACE, on the other hand, the corresponding task is to identify the entity so named. (Doddington et al., 2004, p. 837)

Se no MUC e no CoNLL, o ponto de partida para a identificação de EM era a identificação de nomes próprios – e mesmo assim, principalmente aos substantivos que pertenciam as classes pessoa, local e organização –, no ACE o ponto de partida é o próprio conteúdo. De outro modo, é como se no ACE a tarefa de identificar correferência se misturasse com a tarefa de identificar entidades. Logo, na frase "Quando vinha para casa de táxi, encontrei um amigo e o trouxe até Copacabana; e contei a ele que lá em cima, além das nuvens, estava um luar lindo" (de "A outra noite", crônica de Rubem Braga) os termos 'amigo', 'o' e 'ele' seriam marcáveis como entidades, porque são correferentes, ainda que não sejam nomes próprios. Se substituíssemos o termo "amigo" por Rubem Braga, seriam marcáveis

o nome próprio “Rubem Braga” e os termos “o” e “ele”, pois todas essas palavras se referem a uma mesma entidade.

A perspectiva do ACE sobre o vale como entidade, por um lado, engrandece a noção de entidade, mas, por outro, dificulta bastante a tarefa do ponto de vista da máquina, uma vez que é muito mais fácil partir dos nomes próprios, já que eles, até mesmo pela forma – uso de letra maiúscula – são mais fáceis de serem identificados, enquanto uma noção mais ampliada de entidade engloba conhecimento semântico.

> O ACE, pelo contrário, propôs a pista de EDT - Entity Detection and Tracking, em que o objectivo é fazer o reconhecimento de entidades, quer sejam quer não mencionadas através de um nome próprio, o que alarga consideravelmente a dificuldade da tarefa. O REM passa pois no ACE a compreender todo o reconhecimento semântico de entidades, sejam elas descritas por nomes comuns, próprios, pronomes, ou sintagmas nominais de tamanho considerável. (SANTOS & CARDOSO, 2007, p. 5)

Além disso, no ACE, há um aumento significativo das classes de entidade estabelecidas: pessoa, organização, local, instalação, arma, veículo e entidade geopolítica. Esta última – entidade geopolítica – seria uma categoria feita para englobar entidades ambíguas com relação a serem um lugar ou uma organização. Por exemplo, na frase “Portugal foi o único país da EU que decretou feriado na terça-feira passada”, a palavra Portugal tanto pode ser interpretada como um local, como pode ser entendida como uma organização. Como observou Jiang et al. (2016), a categoria organização se mostra uma das mais difíceis na tarefa de reconhecimento de entidade. Para dar conta de casos menos transparentes, é que uma nova categoria no ACE foi criada.

Outra competição que, assim como o ACE, dá lugar a uma análise mais semântica da entidade é o HAREM (Avaliação e Reconhecimento de Entidades Mencionadas) (SANTOS & CARDOSO, 2007). O HAREM foi organizado pela Linguateca e é a primeira avaliação conjunta de REM para a língua portuguesa. Ele teve duas edições: a primeira ocorreu entre os anos de 2004 a 2006 e a segunda em 2007 e 2008.

De maneira geral, o que contou como entidade, tal como no MUC, foram os nomes próprios. No entanto, enquanto no MUC e no CoNLL, apenas algumas classes interessavam – principalmente as que se referissem a pessoas, locais e organizações –, na HAREM, todos os nomes próprios interessavam. Na realidade,

o HAREM não partiu de categorias pré-definidas de entidades – como a MUC e a CoNLL –, as classes de entidade foram escolhidas depois de uma avaliação textual, em função do que o *corpus* apresentava (SANTOS & CARDOSO, 2007).

No que tange às classes de entidade, as categorias da primeira edição praticamente permaneceram as mesmas na segunda. O HAREM considerou dez categorias de entidade e definiu 41 subtipos de entidade. As categorias utilizadas foram: (1) pessoa; (2) organização; (3) tempo; (4) local; (5) obra (6) acontecimento; (7) abstração; (8) coisa; (9) valor e (10) variado. Em comparação com as primeiras competições descritas, houve um aumento significativo das classes. Inclusive, em comparação com o ACE, também houve aumento, já que o ACE considerou sete tipos.

Considerando ambas as edições, é possível dizer que o HAREM teve três características fundamentais: (1) utilizou um modelo semântico para detecção de EM, o que significa que o valor da entidade estava ligado ao contexto em que aparecesse e não ao significado de um dicionário, (2) trabalhou com a ideia de vagueza, isto é, considerou mais de uma resposta correta para a classificação ou tipo da entidade, e (3) operou com a ideia de flexibilidade, que significou que os participantes não precisavam atuar em todas as tarefas, mas selecionar somente as que interessavam (FREITAS et al., 2010).

A maior contribuição do HAREM, nesse histórico de REM, tem relação com as características um e dois, mencionadas acima. No HAREM, a classificação da entidade leva em consideração o contexto em que ela aparece. Desse modo, a avaliação se torna muito mais difícil. Para ilustrar, a palavra "Brasil" tanto pôde ser classificada como local, ou como organização, ou como pessoa. Em "O Brasil venceu a copa", temos a classe pessoa e o tipo "grupomembro". Já em "O Brasil assinou o tratado", temos a classe organização e o tipo administração. Por fim, em "O Brasil tem muitos rios", temos a classe local e o tipo administrativo.

Além disso, no HAREM, existe a possibilidade de se marcar múltiplas respostas com relação a uma entidade. Ou seja, os sistemas, em casos de vagueza ou ambiguidade, não são obrigados a fazer uma opção, porque o HAREM objetiva trabalhar com todas as perspectivas possíveis pela língua. Por exemplo, na frase "Com a proclamação da carta, temos a obrigação e a oportunidade de dar aos quase 500 milhões de cidadãos e ideia de uma Europa", a entidade "Europa" pode ser classificada como local e pessoa ao mesmo tempo. Não se trata de aceitar as duas

coisas, uma resposta ou outra, mas sim de possibilitar a dupla resposta, sendo uma coisa e a outra também.

No ACE, a marcação múltipla, em até certo ponto, também é possível, já que eles criaram uma classe chamada "entidades geopolíticas" para os casos em que local, pessoa e organização podem ser confundidos, ou funcionarem juntos. No entanto, como explicam Santos & Cardoso (2007), em outras possíveis polissemias, o ACE exigiria uma única resposta correta, enquanto o HAREM possibilita a vagueza de resposta para qualquer caso.

> Para sermos completamente justos, convém realçar, mais uma vez agradecendo à Cristina Mota por nos ter tornado cientes desse facto, que o ACE permite, opcionalmente, a marcação da vertente (local, pessoa, organização) para as entidades geopolíticas. Embora isso seja uma forma de resolver (para um conjunto limitado) a questão das múltiplas vertentes, parece-nos que a diferença é maior que a semelhança: por um lado, no HAREM não é só a categoria <LOCAL|ORGANIZAÇÃO> que pode ser vaga, mas todas; (...) (SANTOS & CARDOSO, 2007, p. 53)

Por outro lado, se uma palavra indica apenas local, a expressão deve ser marcada apenas como local. Não é porque há a possibilidade de marcar mais de uma resposta, que mais de uma resposta está correta e será aceita.

Com a segunda edição do HAREM, alguns atributos novos foram trazidos, são eles: primeiro, a inclusão de duas novas tarefas de PLN junto com REM – reconhecimento de entidades temporais e detecção de relação semântica entre entidades nomeadas, o que incluía correferência, mas não se limitava a isso. E, segundo, quanto às entidades embutidas, o HAREM não considerou apenas uma única alternativa como correta. Ao contrário, como no MUC - através do mecanismo ALT -, a partir da segunda edição, a competição considerou todas as possibilidades como corretas. Por exemplo, o termo "Jogos Olímpicos de Barcelona" podia ser classificado como uma entidade só ou como duas, nesse caso, separando "Barcelona" como entidade específica de lugar (FREITAS et al., 2010).

Atualmente, para encerrar o histórico das competições, é interessante dizer que o trabalho de Yamada et al. (2020) representa o estado da arte na tarefa de REM. Seu modelo foi treinado em diversos *datasets*, inclusive, no ConLL de 2003, no qual obteve o desempenho de 94.3. A abordagem utilizada por Yamada et al. (2020) é o aprendizado profundo (*Deep Learning*). O sistema LUKE (Language Understanding with Knowledge-based Embeddings) trabalha usando

representações de entidades distribuídas como *embeddings*, isto é, o sistema opera criando representações contextuais de entidades. De acordo com os autores, um trabalho futuro é a aplicação do LUKE em tarefas de domínio específico, uma vez que ele operou bem nos *datasets* de linguagem sem domínios especializados.

Para além do MUC, CoNLL, ACE e HAREM, há muitos outros projetos envolvendo o reconhecimento e classificação de entidades mencionadas. A intenção aqui não é fazer uma revisão completa, mas, trazer os trabalhos mais referenciados ou relevantes para a presente pesquisa. O que a breve revisão apresentada buscou mostrar é que o termo "entidade mencionada" – e a tarefa de identificá-la – pode significar/considerar coisas distintas, dependendo, por exemplo, das circunstâncias em que é aplicado e da forma como é avaliado.

Inicialmente, a tarefa começa de forma um pouco ingênua, buscando dar conta de poucas classes de substantivos próprios, típicos de textos jornalísticos. Atualmente, as competições de entidade mais recentes já consideram todos os nomes próprios, e não apenas os de algumas categorias. No panorama histórico traçado nesta seção, um exemplo disso é o HAREM. Além disso, se em competições como MUC, CoNLL e HAREM entidades mencionadas equivaliam apenas a nomes próprios, em projetos como o ACE, elas passaram a serem entendidas de forma ampliada.

Apesar dessa forma mais ampliada de se entender as entidades trazer novos desafios para a tarefa, ela deve ser entendida como um passo fundamental para o aprimoramento de REM, já que, de fato, as unidades de informação relevantes de um texto podem ir muito além de algumas categorias de nomes próprios. Fora dos *corpora* independentes de domínio e dentro de *corpora* de área especializadas ou técnicas – como é o nosso caso – valeria a pena considerar apenas substantivos próprios como entidades? Essa é uma questão que será apresentada e discutida mais adiante.

Não se trata de eleger a melhor ou pior abordagem, mas de tentar dar conta do problema da extração de informação. Se, por um lado, a noção de entidade, como nome próprio, parece ser simplista, com a extensão da noção, novos desafios para a realização da tarefa surgem. Por exemplo, a definição das categorias de entidade passa a requerer uma maior discussão, tal como a classificação das entidades dentro dessas categorias. Isso ocorre porque os nomes próprios possuem, geralmente, uma pista formal na língua, que é a letra maiúscula.

Como pode ser observado através das competições, os tipos de entidade mais comuns são pessoa, local e organização (JIANG et al., 2016). Em *corpora* de domínios/assuntos específicos, no entanto, essas categorias não parecem suficientes para se extrair informação do texto, na medida em que seu conteúdo é mais particular e/ou técnico, especialmente no vocabulário utilizado. Por exemplo, no domínio biomédico, a identificação de doenças ou diagnósticos pode ser mais relevante que a identificação de pessoas, assim como as partes do corpo humano parecem mais relevantes que lugares geográficos. Entretanto, não são todas as doenças e nem as partes do corpo humano nomes próprios iniciados por letra maiúscula. Isso sem contar com todo o vocabulário popular que é utilizado paralelamente aos nomes científicos.

Quando estamos em um domínio específico, a tarefa de identificar entidades mencionadas exige uma metodologia diferente, para além da identificação dos nomes próprios. Isso pode ser observado nos trabalhos de Cohen et al. (2017) e Thompson et al. (2017), sobre *corpora* de domínios biomédicos. Os nomes próprios são bem menos relevantes e informativos nesses domínios, especialmente porque os domínios técnicos possuem uma terminologia muito própria, que dificulta o processamento de informação dos sistemas. A partir da expansão da noção de entidade, de onde partir para identificar e classificar entidades mencionadas em um *corpus* cujo vocabulário é altamente particular? Essa é a pergunta que será respondida ao longo das próximas sessões.

2.4. Recursos lexicais e PLN: taxonomias

Antes de tentarmos oferecer respostas às questões suscitadas nas sessões precedentes, primeiro, devemos (1) pontuar a importância dos recursos linguísticos na Linguística Computacional e (2) delimitar nosso entendimento de taxonomia. Ambas as coisas serão realizadas agora, mas, para tanto, devemos explicar o que são recursos linguísticos e como eles são utilizados, isto é, de que forma são relevantes.

Podemos entender recursos linguísticos como materiais que contêm informações linguísticas. Nesse sentido, *corpus* seria um tipo de recurso. No entanto, no âmbito dessa pesquisa, o que colocamos em voga são os recursos

linguísticos lexicais, isto é, aqueles voltados para palavras e que podem possuir informações como conceitos, definições e relações entre conceitos. De outro modo, enfocamos nos recursos que podem, de maneira ampla, se classificar como léxicos e que podem se apresentar de forma linear, hierárquica ou por redes (FREITAS, 2007).

Os recursos lexicais são materiais que buscam dar alguma organização, ainda que incompleta e não neutra, à língua, ou a uma amostra específica dela. Dentro do PLN, já é reconhecida a importância de tais recursos como: ontologias, taxonomias, tesauros, terminologias, entre outros. A importância se dá, como será mais para frente exemplificado, pelo fato de que recursos lexicais e semânticos podem ser úteis em quase todas as tarefas da Linguística Computacional (KILGARRIFF, 2003).

A ligação entre recursos lexicais e PLN é bem antiga. Ela começa, na década de 70, a partir do momento em que, dentro da área, começa a haver o interesse pela obtenção automática de materiais que pudessem auxiliar as tarefas computacionais. Vale lembrar que ontologias, tesauros e taxonomias durante muitos anos foram elaborados manualmente por lexicógrafos. O interesse por automatizar esse processo objetivava diminuir as dificuldades – e a demora – de se criar, manter e aprimorar os recursos. É a partir do anseio de se obter automaticamente os recursos que começaram a existir diversos estudos voltados para dicionários eletrônicos – MRDs (machine readable dictionaries) (IDE e VÉRONIS, 1995).

Inicialmente, a Linguística Computacional se voltou para a extração de informação em dicionários porque os dicionários eram tidos como recursos de informação tanto lexical quanto semântica. Além disso, os dicionários também eram vistos como uma base de conhecimento confiável e de qualidade para a construção de ontologias lexicais (IDE e VÉRONIS, 1995). As ontologias, ao serem construídas, a partir dos dicionários eletrônicos, ofereceriam informações aos sistemas.

Saindo de um panorama histórico e buscando trazer significados para os recursos, em termos formais, há uma certa imprecisão conceitual entre as noções de ontologia, tesauro, taxonomia, dicionário e todos os outros nomes para os recursos linguísticos. Isto é, é difícil definir e diferir exatamente cada um deles, pois, de uma forma geral, há uma sobreposição inclusive quanto à utilidade: todos facilitam buscas semânticas (FREITAS, 2007). No âmbito no PLN, esses termos muitas

vezes são utilizados sem um acordo sobre suas definições e, por isso, encontram-se na literatura diversas posições.

Se tivéssemos que estabelecer as fronteiras, poderíamos começar pela forma de organização, já que alguns apresentam uma estrutura mais linear, como os dicionários, e, outros, apresentam uma estrutura mais hierárquica, como as taxonomias. Além disso, o nome "léxico" é normalmente utilizado para os recursos cuja natureza do conteúdo não é tão semântica ou não visa uma formalização sobre um conhecimento – como ocorre nas ontologias. O tipo de relação existente em cada recurso certamente pode ser um critério definidor. Se em ontologias e tesauros, diversos tipos de relação – como hiperonímia e sinonímia – podem ser encontrados, em taxonomias encontramos apenas relações hierárquicas (hiperonímias).

> Uma *taxonomia* é uma hierarquia de termos, na qual podem existir diferentes tipos de relação pai-filho (parte-todo; tipo-instância). Já um *tesauro* pode ser considerado uma extensão de taxonomia, comportando a inclusão de regras de uso de vocabulário, definições, sinônimos e antônimos. Compreende, portanto, além de relações hierárquicas, relações associativas. Por fim, *ontologias* (as específicas de domínio, pelo menos) são mais detalhadas; podem – e devem – conter mais níveis hierárquicos. (FREITAS, 2007, p. 35)

A utilização de recursos linguísticos no PLN significa tentar fazer uso de conhecimentos semânticos, lexicais ou gramaticais em tarefas computacionais, isto é, ela implica transformar o conhecimento humano em uma ferramenta cujo formato possa auxiliar a realização de tarefas. Uma forma mais explícita de mostrar como os recursos podem ser utilizados é trazendo exemplos de aplicações. Na tarefa de resolução de correferência/anáfora, por exemplo, a utilização de recursos semânticos frequentemente pode ser observada.

Superficialmente mencionada na seção sobre EM – na parte do projeto ACE –, a correferência é um fenômeno que ocorre quando dois ou mais elementos em um texto se referem a mesma coisa, ou melhor, a mesma entidade (Cohen et al., 2017). Em outras palavras, a correferência significa que palavras diferentes podem ser usadas para se referir a mesma coisa ao longo do discurso. Por exemplo, na frase de Rubem Braga, já mencionada, "Quando vinha para casa de táxi, encontrei um amigo e o trouxe até Copacabana; e contei a ele que lá em cima, além das nuvens, estava um luar lindo" ocorre correferência entre os termos 'amigo', 'o' e 'ele'. Ou seja, essas expressões se referem a uma mesma entidade.

A tarefa de Resolução de Correferência – também chamada de Resolução de Anáfora –, dentro da Linguística Computacional, consiste, para a máquina, na identificação automática dos termos correferentes, isto é, na identificação de palavras que possuem a mesma identidade.

Muitas vezes, a correferência pode ser feita por meio de sinônimos ou hiperônimos. Por exemplo, na frase "Comprei banana na feira, mas a fruta veio estragada", a palavra "fruta" faz correferência a "banana", sendo um hiperônimo dela. Já na frase, "Segure o cão ao abrir a porta, se não o cachorro vai fugir", temos que "cachorro" faz correferência e é sinônimo de "cão". Por conta desta realidade do fenômeno, recursos linguísticos que contenham diversos exemplos de sinônimos e hiperônimos, dentro de um domínio específico ou não, podem auxiliar os sistemas a realizarem a anotação em *corpora*.

O CoNLL, como já vimos, é uma conferência que realiza anualmente competições entre sistemas. Em 2011, com a edição voltada para correferência, foi permitida a utilização da WordNet, tendo em vista a dificuldade da tarefa. A WordNet, no universo do PLN, é o recurso lexical de maior referência. Ela é uma base de dados lexical para o inglês que lembra o formato de um tesauro. Em sua estrutura básica, a WordNet apresenta o *synset* – conjunto de sinônimos para um determinado conceito –, bem como outras informações, como hiperonímia, meronímia, antônimos etc. (GONÇALO OLIVEIRA et al., 2010).

> More so than previous CoNLL tasks, coreference predictions depend on world knowledge, and many state-of-the-art systems use information from external resources such as WordNet, which can add a layer that helps the system to recognize semantic connections between the various lexicalized mentions in the text. (PRADHAN et al., 2011, p. 7)

Por outro lado, no que se refere a identificação e classificação de entidades mencionadas – tema desta dissertação –, na tarefa 7, do SemEval-2018 (uma competição de sistemas, tal como o MUC, porém voltada para questões semânticas), recursos linguísticos foram utilizados como auxílio para a anotação das entidades. Na realidade, a principal tarefa do SemEvel-2018 foi voltada para anotação de relações semânticas, de variados tipos, em *corpora* de domínios especializados. Todavia, para a realização da tarefa principal, foram oferecidos *corpora* anotados com entidade, tanto no material de teste, como no de treino (GÁBOR et al., 2018).

A anotação das entidades não foi realizada por meio da identificação de nomes próprios de categorias pré-definidas, mas sim por um anotador automático, que, por sua vez, era baseado em um almanaque, feito a partir de ferramentas de extração de terminologia e recursos ontológicos. Dessa forma, as entidades anotadas eram palavras diversas do domínio científico. Com isso, os sistemas participantes não precisaram anotar relações semânticas entre qualquer par de palavras no *corpus*, mas sim entre as entidades previamente anotadas. Cabe destacar que a tarefa de classificação da relação foi dividida em duas partes: em uma, os sistemas lidaram com um *corpus* sem revisão humana, na outra, os sistemas atuaram em cima de um *corpus* revisado (GÁBOR et al., 2018).

Até aqui, buscamos brevemente exemplificar como esses recursos lexicais são utilizados e de que forma são relevantes. Tais recursos nada mais são do que uma tentativa de organizar o conhecimento do mundo ou uma língua. No caso de recursos de domínios específicos, eles procuram dar conta do conhecimento de uma área, ou de uma língua de especialidade, ou ainda de todo um modo de expressão particular (KRIEGER e FINATTO, 2004).

Uma vez que o objetivo deste trabalho é apresentar uma taxonomia do domínio de óleo e gás, cabe, no final da seção, uma elucidação mais específica do que se subentende neste trabalho pela noção de taxonomia. Assumindo a definição trazida em Freitas (2007), no escopo desta pesquisa, taxonomias devem ser compreendidas, de forma simplificada, como uma terminologia estruturada, com relações de hiperonímia. De modo geral, as taxonomias podem ser utilizadas, tal como as ontologias. Elas podem servir no oferecimento de relações de hiperonímia entre conceitos e podem oferecer um vocabulário especializado.

> It's not surprising, therefore, that identifying hypernymic (is-a) relations has been pursued in NLP for more than two decades (Shwartz et al., 2016): indeed, successfully identifying this lexical relation substancially improves Question Answering applications (Prager et al., 2008; Yahya et al., 2013) Textual Entailment and Semantic Search systems (Hoffart et al., 2014; Holler et al., 2014; Roller and Erk, 2016). In addition, hypernymic relations are the backbone of almost every ontology, semantic network and taxonomy (Yu et al., 2015), which are in turn useful resources for downstream tasks such as web retrieval, website navigation or records management (Bordea et al., 2015). (CAMACHO-COLLADOS et al., 2018, p. 712)

No âmbito deste trabalho, o que se espera é que a taxonomia criada possa ser utilizada como subsídio em tarefas da Linguística Computacional,

especialmente, na tarefa de identificação e classificação de entidades mencionadas. Nessa seção, através da descrição do SemEval-2018, sem tanto aprofundamento, já foi possível observar que um recurso linguístico – construído com base em terminologia – pode ser uma maneira paralela de se anotar entidades em um *corpus*. Ao longo das próximas seções, ficará ainda mais claro como uma taxonomia da área de óleo e gás pode servir na tarefa de REM, dentro de um *corpus* do domínio de óleo e gás, mas, para chegarmos lá, uma breve contextualização acerca da relação entre "termos" e "entidades" ainda se faz necessária.

2.5. Termo e Entidade Mencionada: uma aproximação

Nesta pesquisa, partimos de glossários e dicionários da área de óleo e gás para construir uma taxonomia que possa ser utilizada como recurso em tarefas do PLN. Vale mencionarmos que os glossários e dicionários são entendidos como terminologias que compreendem os termos de uma área. De igual maneira, glossários/dicionários/tesauros também podem ser entendidos como a face prática da Terminologia, como área. Desse modo, produzir glossários é produzir instrumentos de organização formal das terminologias. Se partimos de glossários e dicionários da área de óleo e gás para construir uma taxonomia, podemos dizer que, por mais redundante que seja, partimos de terminologias para a construção da taxonomia.

É nisso que se compreende a importância das terminologias para o PLN, por oferecerem uma organização de uma língua de especificidade, elas podem ser utilizadas na construção de recursos linguísticos de domínio, que, por sua vez, podem ser utilizados em tarefas da Linguística Computacional. É exatamente isso que ocorreu no SemEval-2018, descrito anteriormente, as entidades do *corpus* foram anotadas com base em um recurso lexical que, por sua vez, foi feito por meio de terminologias.

> Specialist domains are characterised by extensive use of technical and domain specific terminology. Term recognition is an important step towards Named Entity Recognition (NER) in these domains: entities, or things in the real world, are often referred to by terms in the text. Large scale knowledge resources such as terminologies and ontologies are typically available in these same domains. We might expect such resources to have some use in term and entity recognition. We

> might also expect entity recognition to add value by linking entities back to these knowledge resources, making additional information available to applications and their users. (ROBERTS et al., 2008, p. 2974)

Como explica Roberts et al. (2008), terminologias são recursos que podem ser utilizados na tarefa de identificação de entidade mencionadas. Especialmente no contexto de entidades de domínios técnicos ou específicos, as terminologias podem ser um ponto de partida, porque oferecem léxicos especializados de uma área, isto é, elas apresentam os termos que, a princípio, seriam relevantes de um dado domínio.

Drouin (2003), pesquisador que desenvolveu uma ferramenta[*] de extração automática de termos a partir de *corpus*, afirma que o *corpus* técnico possui itens lexicais que estão relacionados ao objeto de uma área, isto é, a frequência de palavras e sintagmas de domínio específico são maiores nesse tipo de coleção de texto do que em outros *corpora*. Por conta disso, para o autor, o *corpus* pode ajudar a se ter acesso a uma determinada terminologia, ou melhor, o *corpus* pode ser o recurso com o qual se constrói uma terminologia.

De modo complementar, e em concordância, afirmo que a terminologia também pode ser o recurso que oferece candidatos a entidade mencionada em *corpora*. Isto é, os termos de terminologias de domínio são os candidatos a entidade em *corpora* de domínio. O que quer dizer que, se extrair entidades mencionadas de domínio significa extrair uma terminologia, construir uma terminologia de domínio é oferecer candidatos a entidade mencionada para *corpora*.

Na Linguística Computacional, o termo – da Terminologia – acaba sendo a entidade mencionada, nos domínios especializados. Isto porque chamamos de entidade todos os itens lexicais relevantes que aparecem dentro de um *corpus*. As entidades são unidades de informação que oferecem dados importantes sobre o que está sendo dito em um *corpus*/domínio. *Corpora* mais gerais possuem entidades mais gerais, mas *corpora* de domínios específicos possuem entidades específicas, que são os termos relevantes da área.

Creio que isto responda às indagações deixadas nas seções anteriores, sobre os desafios de se identificar entidades no domínio técnico. Quando estamos em um domínio específico, a tarefa de identificar entidades mencionadas exige uma

[*] Drouin (2003) desenvolveu a ferramenta TermoStat, que, inclusive foi utilizada pela pesquisa. Não nos aprofundamos nisso agora, porque mais para frente será detalhado.

metodologia especial, não bastando identificar os nomes próprios, é preciso reconhecer os nomes, as palavras, os termos relevantes dentro daquele contexto. E isso a terminologia oferece. Ao mesmo tempo, as terminologias, assim como os *corpora*, podem ser usadas para a construção de recursos linguísticos com ainda outros diversos tipos de informação, que, por sua vez, podem ser úteis em diversas tarefas do PLN, tal como na classificação de entidades e na identificação de relação entre entidades.

Outro objetivo dessa dissertação é estabelecer um cruzamento entre Linguística (Terminologia) e PLN (entidades mencionadas). Por conta disto, é imprescindível finalizar esta seção apontando, como linguista, para o fato de que há, nitidamente, uma articulação, prática e teórica, entre uma área aplicada dos estudos da linguagem e uma aplicação de extrema relevância do PLN. Apesar de que essa relação não é vista – pelo menos com frequência – formalmente apresentada, em ambas as áreas, ela não é exatamente uma novidade. Isto é, em diversos trabalhos, como os mencionados na dissertação, por exemplo, é possível perceber o cruzamento, ainda que sem tanto diálogo.

2.6. Entidades Mencionadas no domínio técnico

Tendo em mente que o domínio técnico exige uma metodologia específica para identificação de entidades e que um bom ponto de partida são as terminologias da área, como mostrado até aqui, evidenciamos agora duas etapas fundamentais para a obtenção de entidades em *corpora* técnicos: (1) construção de uma organização do vocabulário da área – seja uma terminologia, uma taxonomia ou uma ontologia – e (2) verificação da cobertura da organização no *corpus* e verificação das instâncias em que os candidatos aparecem.

Como este trabalho não chega de fato a fazer a anotação de EM em um *corpus* – mas sim a apresentar os primeiros subsídios para que isso seja feito na área de óleo e gás –, focamos, nessa seção, na apresentação de procedimentos que envolvem a primeira etapa, bem como questões metodológicas sobre eles. Então, para organizar o vocabulário de uma área, pode-se partir de diversas fontes, dentre elas: (1) vocabulários especializados públicos; (2) um *corpus* da área; e (3) metadados de um *corpus* da área.

Sobre o primeiro tipo, ele já foi justificado na seção anterior. Uma vez que extrair as entidades de um domínio também pode querer dizer extrair uma terminologia (e vice-versa), se existem terminologias prontas (dicionários e glossários especializados, por exemplo), elas servirão para apontar as palavras relevantes da área e, sobretudo, possíveis entidades, bem como relações entre elas. Logo, a utilização de glossários e dicionários especializados, públicos e disponíveis, pode ser um ponto de partida para se conhecer os vocábulos de um domínio. Além disso, outros recursos, como tesauros, também são bem-vindos.

É possível encontrar, na literatura disponível, autores que aproximam, utilizando como sinônimos, as palavras "termo", oriundo das terminologias, e "descritores", que são listados nos tesauros (VAN DER LAAN, 2002). Isso, a princípio, parece demonstrar que os tesauros – se não apresentam os termos de um domínio –, pelo menos, apresentam sintagmas relevantes na área, com potencial de serem termos. Desse modo, partindo do objetivo de se organizar um vocabulário de domínio que sirva de pista para o reconhecimento de entidade mencionada em *corpora*, os tesauros, bem outros recursos lexicais, constituem uma possível fonte.

Uma diferença entre terminologias e tesauros, no entanto, pode ser colocada. Os tesauros funcionam como um vocabulário que busca controlar a terminologia de um domínio. Isto é, eles apresentam diversas relações entre os termos, que visam esclarecer sobre possíveis ambiguidades e polissemias. Por conta disso, os descritores possuem também uma natureza muito mais pragmática que os termos.

> Entendemos os tesauros como a relação de termos de uma linguagem de especialidade, estruturada de modo a evidenciar as relações conceituais dessa área do conhecimento. Dessa forma, estamos afirmando que os descritores serão melhor definidos se tratados na perspectiva de unidades lexicais especializadas utilizadas no discurso de uma determinada área do conhecimento (...) Por sua estrutura, os tesauros, hoje, representam um dos instrumentos de controle de vocabulário mais utilizados em sistemas de informação. (VAN DER LAAN, 2002, p. 18).

Todavia, para além de terminologias disponíveis, já existe também, com foi ligeiramente comentado, a possibilidade de se extrair automaticamente uma terminologia, através de *corpus*. Logo, quanto aos *corpora* – por serem um conjunto de textos sobre o assunto do qual se deseja obter entidades/termos –, eles também podem servir de base para se obter entidades/termos. Isso significa que podemos identificar as entidades de um *corpus* utilizando o próprio *corpus* como fonte de

candidatas. Isso geralmente é feito por meio de uma ferramenta capaz de anotar as classes gramaticais e funções sintáticas de todas as palavras de um *corpus*. A partir dessa anotação, é possível reconhecer e identificar – por métodos estatísticos – quais sintagmas nominais são mais frequentes, específicos e relevantes de um domínio em particular (KRIEGER & FINATTO, 2004).

Quanto a este processo, ele implica a utilização de procedimentos automáticos baseados em abordagens estatísticas, linguísticas e/ou híbridas, para extração de termos ou sintagmas relevantes. Cabe, ainda nesta seção, explicitar melhor algumas características dessas abordagens.

Na abordagem estatística, em primeiro lugar, o processo se baseia em frequência de ocorrência: "os documentos contidos no *corpus* são vistos como um conjunto de termos e é medida sua frequência de ocorrência" (Lopes et al., 2009, p. 79). Os programas normalmente identificam alguns ngrams de um *corpus* e fazem isso seguindo alguns parâmetros: (1) determinam o tamanho dos ngrams; (2) usam uma *stoplist*, que contém e exclui as categorias morfológicas sem valor terminológico e palavras não relevantes; (3) determinam um limite de corte, que indica valores de frequência que devem ser eliminados e (4) aplicam regras de formação de palavra, que explicam o padrão de palavras que deve ser selecionado, por exemplo, se letras maiúsculas e minúsculas são aceitas. Um exemplo de ferramenta que trabalha exclusivamente com o método estatístico na identificação de palavras-chave é o AntConc (Anthony, 2005), que será utilizado nessa pesquisa.

Já a abordagem linguística está relacionada a um método baseado em regras linguísticas. Nesse caso, todo o processo começa e depende da anotação linguística do *corpus* (Lopes et al., 2009; Drouin, 2003). A partir do *corpus* anotado, a ferramenta extrai sintagmas nominais de tamanhos e combinações morfológicas distintos. Os sintagmas extraídos passam por algumas regras linguísticas que servem para limpar as entradas de unidades lexicais desnecessárias.

Tanto os estilos de anotação quanto as regras linguísticas utilizadas variam de trabalho para trabalho. Ao final, os itens são geralmente apresentados seguindo um critério de frequência. Um exemplo de trabalho para a língua portuguesa e para um domínio próximo, e com ênfase neste método, é o de Wendt et al. (2010), que utiliza a ferramenta ExatoLP para construir um glossário na área de geologia. Ao mesmo tempo que a ênfase é na linguística, vale lembrar que essa pesquisa também faz computação de frequência, logo, a abordagem não é exclusivamente linguística.

Por fim, a abordagem híbrida é a combinação das duas abordagens, estatística e linguística. Um exemplo de trabalho híbrido é o de Lopes et al. (2009), no domínio pediátrico. Os autores começam a extração de sintagmas pela anotação linguística, mas, posteriormente, empregam diversos métodos de estatística e frequência para extrair sintagmas simples e compostos. Outro exemplo é o trabalho de Park et al. (2002), que também parte da anotação de POS para, em seguida, usar cálculos estatísticos para filtrar pré-modificadores genéricos e ranquear os resultados obtidos. Na próxima seção, estes trabalhos serão explicados de maneira mais detalhada.

Entre os três tipos de abordagem, qual seria o melhor? E quais as vantagens e desvantagens de cada método? Lopes et al. (2010) fizeram uma pesquisa comparativa utilizando ferramentas com abordagens distintas. Utilizaram, em um mesmo experimento, a ExatoLP, com um método mais linguístico, e a NSP, com um método mais estatístico. Os resultados mostraram que a abordagem estatística obtém resultados melhores na abrangência, enquanto a abordagem linguística tem um desempenho melhor na precisão. Ou seja, o primeiro método extrai mais entradas que o segundo, mas o segundo, apesar de extrair menos, extrai boas entradas.

Por fim, além da extração automática – via o conteúdo do *corpus* – de um vocabulário de domínio, também é possível partir da ideia de que não só o *corpus* pode ser utilizado, mas também os metadados do *corpus*. Por exemplo, em um corpus de textos acadêmicos, as palavras-chave dos textos – dissertações, teses, monografias, artigos –, quando armazenadas/extraídas separadamente, servem para se obter uma lista de possíveis entidades, com termos ou sintagmas relevantes indicados por pesquisadores da área. Uma vez que as palavras-chave representam os termos relevantes dos trabalhos acadêmicos que compõem um *corpus* e que foram escritas manualmente por, a priori, especialistas da área, elas contam como uma espécie de terminologia, de vocabulário, da área também. Isto é, elas também são unidades linguísticas que funcionam como pistas sobre o assunto que um corpus de textos acadêmicos aborda.

No que concerne à metodologia deste trabalho, utilizamos três tipos de fonte – vocabulários (terminologias e tesauros), *corpus* e metadados (palavras-chave). Além disso, com relação aos procedimentos no *corpus*, não visamos a utilização exclusiva de um único método porque partimos do pressuposto de que nenhuma

abordagem é completa e de que a complementação de estratégias é sempre bem-vinda. Por isso, uma abordagem híbrida foi utilizada para extrair possíveis termos de O&G do *corpus* Petrolês. Após obtermos candidatos a termos da área, elaboramos uma taxonomia, com todas as entradas obtidas, vindas das diversas fontes. A taxonomia elaborada é o recurso que propomos para auxiliar na identificação entidades do domínio. No capítulo 4, todos os procedimentos aqui brevemente mencionados serão descritos com mais detalhes. Entretanto, antes, faremos uma revisão bibliográfica dos trabalhos que serviram de referência metodológica nesta tarefa.

3
Trabalhos relacionados

Esta seção tem o objetivo de descrever trabalhos que inspiraram a presente pesquisa. Nas subseções, inicialmente, mencionamos autores que partiram de *corpus* para se obter terminologias de domínio. Tais autores dialogam com esta pesquisa, na medida em que ela se propõe, partindo de um *corpus* de óleo e gás, a construir uma taxonomia – derivada de terminologias – que auxilie em tarefas do PLN, especialmente na tarefa de identificação e classificação de entidades mencionadas. Em seguida, trazemos trabalhos sobre geração automática de taxonomias. Nessa seção, interessa justamente destacar como, partindo de *corpus* e de terminologias (como é o nosso caso), são elaboradas as taxonomias. Ao final, detalhamos dois *corpora* de domínio técnico, já mencionados anteriormente, que inspiraram esta dissertação: GENIA e CRAFT.

Antes de propriamente chegarmos às subseções, no entanto, é válido mencionarmos, mais uma vez, trabalho de Cleverley e Burnett (2015), uma vez que esses autores acabaram sendo uma grande motivação para a realização desta pesquisa, dentro do universo do petróleo.

Sobre o estudo de caso realizado por Cleverley e Burnett (2015), eles pontuam três questões de pesquisa para serem exploradas no experimento: (1) até que ponto um tesauro pode ser aprimorado por meio de técnicas automatizadas, (2) qual é o valor do conteúdo de categorização automática que já foi classificado manualmente e (3) até que ponto as técnicas de KOS (Sistemas de Organização do Conhecimento) manuais e automáticas podem ser combinadas em uma interface de usuário de pesquisa para estimular descobertas?

A análise foi realizada em uma grande empresa de petróleo e gás. Foram recrutados seis geólogos para ajudar a responder à pergunta 2 e dezesseis geofísicos para a pergunta 3. Doze geocientistas (dois grupos de seis) forneceram informações adicionais. As questões de pesquisa 1 e 2 foram abordadas paralelamente, considerando um problema real de negócios identificado na empresa.

Na empresa, o caso era o seguinte: uma equipe de exploração de petróleo e gás tinha em torno de 13.000 documentos eletrônicos de escritório em seu sistema

de arquivos compartilhado. No entanto, eles estavam organizados em pastas, o que só permitia pesquisas feitas por nome de arquivo. De acordo com os autores, era fácil para a equipe perder informações porque nem todos os tópicos dos documentos apareciam no nome do arquivo, tampouco no nome das pastas. Além disso, havia membros de equipe que não estavam familiarizados com o conteúdo e organização das pastas.

O estudo de caso visou melhorar esse contexto específico da empresa. Para tanto, os 13.000 arquivos foram indexados em um texto completo – isto é passaram a ser, de certo modo, um *corpus,* uma coletânea de vários textos. Em seguida, com base em um tesauro da área de óleo e gás – elaborado manualmente –, os documentos foram categorizados/anotados automaticamente com as entidades e os termos sinônimos. Uma vez que a sinonímia anotada levou em conta as relações do tesauro construído e dos termos presentes nos documentos, o próprio recurso linguístico – o tesauro – pôde ser aprimorado.

Sobre os resultados da questão 1 – até que ponto um tesauro da área de O&G pode ser aprimorado por meio de técnicas automatizadas – os dados se pautaram no aprimoramento do tesauro que havia sido construído manualmente e, que, posteriormente, foi apurado devido à anotação do *corpus*. Os resultados mostraram que a combinação de técnicas, automáticas e manuais, oferece um nível de qualidade muito maior do que a utilização de um único método.

Sobre os resultados que envolvem a questão 2 (qual é o valor de um conteúdo, já classificado manualmente, que foi categorizado automaticamente), a análise recaiu sobre a mudança de organização dos 13 mil arquivos da empresa. Os participantes concordaram de forma unânime que pesquisar os documentos em texto completo, por *corpus*, e navegar por facetas categorizadas automaticamente, ao invés das pastas, melhorou o desempenho em encontrar e descobrir informações em grande escala.

Quanto à questão 3 – sobre até que ponto as técnicas de KOS manuais e automatizadas podem ser combinadas em uma interface de usuário de pesquisa para estimular o "acaso" –, é válido dizer, antes de mais nada, que essa questão tem a ver oferecer possibilidades aos participantes de encontrarem informações pelas quais não esperam ou não exatamente buscam. Para tanto, foi criado um "estimulante" semi-interativo, que foi projetado para provocar interação e

discussão, usando 70.000 resumos de artigos, vindos da *Society of Petroleum Engineers* (SPE).

Dezesseis geofísicos participaram desse estudo. A interface foi disponibilizada em grandes telas (*touchscreen*) e as interações dos participantes foram gravadas em vídeo. Cada sessão durou 45 minutos e teve de dois a nove funcionários. Sobre os resultados, os participantes afirmaram que a proposta seria poderosa para o encontro "ao acaso" de informações e que poderia ajudar na elaboração de taxonomias.

É bom lembrar que a esta dissertação, ao oferecer uma taxonomia de domínio, visa ser uma contribuição prática dentro desse contexto apresentado por Cleverley e Burnett (2015). A partir de uma taxonomia de óleo e gás em português, experimentos semelhantes aos dos autores são possíveis, viabilizando uma melhora na extração e organização de informações dentro da área.

3.1. Extração automática de terminologia em *corpus*

Nesta seção, falaremos sobre trabalhos que lidam com obtenção automática de terminologia – ou, de outro modo, obtenção de candidatos a entidades em domínios técnicos – por meio de *corpus*. Podemos começar mencionando o trabalho de Wendt et al. (2010), cuja finalidade foi construir automaticamente um glossário, através de um *corpus* de geologia. Este trabalho converge com o nosso na medida em que ele visa construir um glossário por meio de *corpus*, o que quer dizer que uma lista de referência no domínio – uma terminologia – será criada.

Pela ótica de Wendt et al. (2010), a tarefa de extração de glossários objetiva, em sua base, organizar palavras e sintagmas de uma coleção de documentos – que são relevantes para um domínio – em uma lista, "itens de um glossário". Pela perspectiva do nosso trabalho, vale lembrar, uma lista gerada automaticamente a partir de um *corpus* pode se tornar também uma lista de candidatas a entidade mencionada em um *corpus*, desde que recebe algum tipo de tratamento.

De forma breve, o trabalho de Wendt et al. (2010) se resume com as duas seguintes etapas: primeiro, extração, de dentro do *corpus* sintaticamente anotado, de sintagmas relevantes para o domínio de geologia. Segundo, recuperação de definições, através de textos da web, para estes sintagmas. As definições foram

recuperadas por meio de diversas fontes, desde glossários específicos, desenvolvidos manualmente por especialistas do domínio, a definições da Wikipédia, ou seja, por fontes de diferentes níveis de confiabilidade.

Quanto à primeira etapa, sintagmas nominais foram extraídos/selecionados por meio da ferramenta ExATOlp (Extrator Automático de Termos para Ontologias em Língua Portuguesa), havendo critérios estatísticos para decidir os melhores candidatos e havendo consideração de informações linguísticas. Foi gerada uma lista com 4626 termos, sendo 1653 unigramas (1 token), 2003 bigramas (2 tokens) e 950 trigramas (3 tokens). Não foram considerados termos com mais de quatro tokens.

Algumas estratégias para extrair apenas sintagmas relevantes foram: (1) são eliminados termos extraídos como SN, mas que terminam com preposição. Por exemplo, "rocha acrescida de", "dosagem diária para"; (2) são eliminados SNs que possuem números, exemplos, "década de 50", "dois estudos"; (3) são excluídos os SNs cujo núcleo não é substantivo, nem nome próprio, nem adjetivo, exemplo, "observado por outros"; (4) SNs que começam com artigos são armazenados sem a primeira palavra (o artigo), por exemplo, o sintagma "a rocha magmática" é armazenado apenas como "rocha magmática".

Essas estratégias no trabalho de Wendt et al. (2010) são interessantes para gerar melhores resultados e eliminar erros. Também podemos constatar algo semelhante em Freitas (2007), que, para construir automaticamente uma ontologia no domínio da saúde, utilizou filtros para eliminar candidatos ruins. Em Freitas (2007), os filtros eliminavam sintagmas com substantivos muitos gerais (exemplo: osteoporose < *fatores*) ou eliminavam sintagmas com adjetivos pré-nominais ou muitos gerais (exemplos: *baixo* rendimento escolar e colesterol *alto*) e pronomes dêiticos (exemplo: broncodilatadores < medicamentos prescritos por *seu* médico).

Quanto à segunda etapa, uma base de definições foi construída com verbetes de dois glossários – UNB e Mincropar – e da Wikipédia e, em seguida, a melhor definição para cada termo foi eleita. Como os dois glossários públicos do domínio não contiveram definições para todos os termos extraídos pelo ExATOlp, definições da Wikipédia foram incluídas. Além disso, sintagmas extraídos que não possuem definições em nenhuma dessas três fontes foram descartados, sendo que um pouco mais da metade dos termos extraídos não foi encontrada nem nos glossários nem na Wikipédia. Termos compostos foram a maioria desses casos.

No fim, sem as definições da Wikipédia, os resultados foram: 926 unigramas, 458 bigramas e 142 trigramas, com precisão de 55%, 23% e 15%, respectivamente e abrangência de 33%, 40% e 32%, respectivamente. Com a inclusão das definições da Wikipédia, o trabalho obteve: 1.367 unigramas – com precisão 82% –, 488 bigramas – com precisão 24% – e 268 trigramas – com precisão 28%.

No que concerne à ferramenta utilizada, a ExATOlp é uma ferramenta baseada em métodos linguísticos e estatísticos e que serve para extrair de um *corpus* seus termos significantes. Além disso, ela também pode computar índices numéricos, auxiliar uma análise comparativa e pesquisar por termos dentro do *corpus* (LOPES et al., 2009).

A ExATOlp funciona da seguinte forma: primeiro, sintagmas nominais – apenas nominais – são extraídos de um *corpus* sintaticamente anotado. Nesse caso específico, o trabalho contou com anotação prévia feita pelo analisador PALAVRAS (Bick, 2000). Assim que a ferramenta detecta todos os sintagmas nominais do *corpus*, os resultados passam por regras de descarte, que servem para eliminar resultados ruins. Essas regras incluem excluir sintagmas (1) com numerais e (2) caracteres especiais. Depois disso, os resultados passam por regras de transformação, que servem para eliminar partes do sintagma, a fim de deixar apenas o mais relevante. São três as regras e transformação: (1) remoção de coordenação no final de um SN. Exemplo: “placas convergentes e divergentes” > placas convergentes. (2) Remoção de pronome no início do SN. Exemplo: “aquelas crianças recém-nascidas” > crianças recém-nascidas. (3) Remoção do artigo, independentemente da posição. Exemplo: “deriva dos continentes” > deriva de continentes. Ao final, o primeiro resultado apresenta sintagmas nominais, compostos pelas anotações: word (forma do sintagma sem alteração), lemma (sintagma na forma canônica), head (núcleo do sintagma), pos (classes gramaticais das palavras do sintagma) e sem (informação semântica).

Em seguida, os sintagmas extraídos passam por uma computação de frequência. Esse processo significa, de forma simples, a contagem de quantos termos foram extraídos e quais as frequências absoluta e relativa para cada um deles. Nessa etapa, os sintagmas também são classificados de acordo com o número de palavras (unigramas, bigramas, etc.) e é possível, caso desejado, eliminar termos menos frequentes.

Ao final desta fase, a ferramenta apresenta: a) frequência absoluta, b) frequência relativa, c) número absoluto de termos e d) percentil dos termos, sendo que, as opções (c) e (d) possibilitam tolerância de resultados. Por exemplo, no caso de um pedido dos 100 termos mais frequentes, o resultado possibilitaria uma lista com mais de 100 itens, caso alguns termos tivessem a mesma frequência. O usuário pode escolher fazer uso desta tolerância ou não.

Na última etapa, os dados chegam ao pós-tratamento. Aqui, a lista pode ser salva e extraída em diversos formatos. É possível extrair uma lista mais simples, com os resultados obtidos, até então, e também é possível comparar os termos extraídos com outras listas de referência no domínio.

No fim, a ferramenta apresenta uma lista de termos relevantes de uma determinada área, que pode ser usada para construir glossários, ontologias, entre outras coisas. Como pontos a melhorar na ferramenta, os autores citam a possibilidade de conseguir lidar com diversos tipos de anotação (não apenas a do PALAVRAS, em xml), e o fato de a regra de remoção da conjunção eliminar resultados interessantes que derivam de coordenação. Exemplo: "placas convergentes e divergentes" poderia abrir dois resultados: "placas convergentes" e "placas divergentes", ao invés de eliminar o que vem como coordenação.

Outro trabalho voltado para extração automática de um glossário de domínio, assim como o de Wendt et al. (2010), é o de PARK et al. (2002). Nesse caso, os autores descrevem o método/procedimento utilizado e apresentam a ferramenta GlossEx, junto com a avaliação de sua performance. Sobre o processo do trabalho, de forma resumida, a extração automática em sua primeira etapa extrai itens candidatos/relevantes em um *corpus* de um domínio específico. Em seguida, esses itens candidatos são revistos, modificados e/ou aprovados por especialistas.

Como dito acima, a primeira etapa da extração do glossário consiste em extrair automaticamente itens candidatos em um *corpus*, isto implica identificar as formas dos itens candidatos através de estruturas sintáticas. Nesta fase de selecionar formas candidatas, foram excluídos: nomes de pessoas e de lugares; sintagmas com mais de seis palavras e tokens especiais que contivessem links ou caracteres especiais, com exceção de hífen e travessão.

A segunda etapa consiste em filtrar os pré-modificadores genéricos, isto é, os que não representam informação relevante para o domínio do glossário. No caso

da referida pesquisa, os pré-modificadores foram decididos por meio de probabilidade e cálculos estatísticos.

Em seguida, na terceira etapa, todas as formas variantes de um conceito foram agregadas em uma forma só, a forma canônica, isto é, todas as variantes se tornam uma única entrada no glossário. O trabalho menciona cinco tipos de variantes: primeiro, as variantes simbólicas, que ocorrem quando termos compostos são separados por diferentes caracteres. Exemplo: *audio/visual input* e *audio-visual input*. Segundo, variações nos compostos, que ocorrem quando há diferenças na escrita de nomes compostos. Exemplo: *Passenger Airbag* e *passenger air bag*. Nesses dois tipos, a forma mais frequente é eleita a canônica. Terceiro, as variantes flexionais, que ocorrem quando o termo flexiona, ao invés de permanecer na forma canônica. Exemplo: *rewinds*, *rewinding*, *rewound*, ao invés de *rewind*. Esta é a variante mais comum. Quarto, as variações de erro ortográfico, que lidam com as palavras escritas de forma errada. Essas palavras são consideradas variantes da forma de dicionário. O quinto e último tipo são as abreviações. Todas as formas abreviadas são variantes da forma por extenso.

Por fim, é utilizada informação estatística para ranquear os itens candidatos obtidos. Cada termo é avaliado em quanto ele e está relacionado com o domínio, para isso é feita uma pesquisa comparativa entre a frequência do termo em um *corpus* geral e um *corpus* de domínio. Além disso, os autores criaram uma própria medida para computar a coesão de termos compostos e compararam com o "coeficiente Dice", um cálculo tradicional.

Algo interessante na pesquisa de Park et al. (2002) é que não apenas sintagmas compostos foram considerados itens candidatos, mas verbos e substantivos simples também, o que não é tão comum. Os possíveis itens de um glossário nomeiam e descrevem conceitos de um domínio, mas eles podem aparecer em sua forma canônica ou como uma variante do conceito. No caso dos verbos, por exemplo, apenas a forma básica do verbo entrava para a lista e os verbos não podiam ser auxiliares. Quanto à estrutura dos sintagmas nominais, eles podem conter determinantes, adjetivos, formas predicativas, conjunções e substantivos.

Agora, mais um trabalho que dialoga com o nosso é o de Lopes et al. (2009) cujo intuito é desenvolver uma ontologia no domínio médico por meio de *corpus* e utilizando uma ferramenta para extração de sintagmas compostos: a OntoLP. Segundo os autores, as ontologias são construídas automaticamente através de cinco

etapas: (1) descoberta de axiomas – extração de termos candidatos a um conceito em um domínio; (2) identificação de instâncias – relações hierárquicas e não-hierárquica entre os termos; (3) determinação de relações; (4) definição de hierarquia de conceitos; e (5) extração de termos candidatos a conceitos. O artigo de Lopes et al. (2009) tem a ênfase no processo e nos resultados da primeira etapa deste processo.

No contexto da OntoLP, a abordagem é híbrida. A extração de termos ocorre em duas fases: CorpusXCES e extração. Na primeira fase é feita a anotação automática do *corpus* através do parser PALAVRAS e, em seguida, se obtém o *corpus* anotado no formato XCES/PLN-BR. As informações linguísticas empregadas são categorias gramaticais, semânticas e de identificação de grupos gramaticais nominais, todas fruto da análise do PALAVRAS. Na segunda fase, diferentes métodos são utilizados para a extração de termos simples e compostos, desde abordagens estatísticas a abordagens linguísticas.

A interface de extração possui três partes: (1) seleção de grupos semânticos; (2) extração de termos simples; e (3) extração de termos compostos. A primeira parte é opcional e consiste no fato de que o usuário pode escolher ou excluir um grupo semântico específico para o trabalho. Por exemplo, ao escolher o grupo "an" (anatomia), palavras anotadas com essa tag serão mostradas, tal como tórax. A segunda parte tem relação com a extração de unigramas, sintagmas simples. E na terceira parte, são extraídos sintagmas compostos, especificamente bigramas e trigramas, que são o foco da pesquisa.

A extração dos termos simples é realizada com base em classes gramaticais. Já a extração de termos compostos, parte tanto de abordagem estatística quanto linguística. É medida, primeiro, a frequência de coocorrência de n-gramas. Por esse método, os autores também trabalham com alguns filtros, como eliminação de candidatos com preposição no início e/ou no fim. Em seguida, é utilizado um método que considera classe gramatical – isto é, a extração é feita por padrões linguísticos. Por fim, com uma abordagem sintática, são extraídos os sintagmas nominais do *corpus*, que foram anotados pelo parser.

Com relação ao experimento, o *corpus* do domínio médico utilizado continha 283 textos do Jornal de Pediatria, num total de 785.448 palavras. Esse *corpus* foi anotado pelo PALAVRAS e recebeu o formato compatível com o OntoLP. Na segunda etapa, grupos semânticos considerados não relevantes foram

excluídos da análise, a extração dos termos simples foi realizada e, em seguida, foi a feita a extração dos termos compostos, nos três métodos: n-gramas, padrões morfossintáticos e o sintagma nominal. Para fins de comparação, o mesmo procedimento foi executado com o *corpus* sem a exclusão de grupos semânticos.

As listas resultantes do experimento foram comparadas com as listas de referência construídas manualmente pelo Grupo TEXTQUIM/TERMISUL, da Universidade Federal do Rio Grande do Sul. A partir desta tarefa de elaboração de material de referência, foram gerados (1) um glossário para apoio a estudantes de tradução e (2) um catálogo de expressões recorrentes em pediatria.

As listas de referência, manualmente elaboradas, também passaram por filtragem automática. Além de eliminação de preposições, também foram eliminados n-gramas contidos em outros n-gramas, ficando somente os n-gramas maiores. A lista gerada foi revisada manualmente por estudantes de tradução e, ao final, o resultado foi de 22.407 termos, sendo 1.293 bigramas, 775 trigramas e 339 termos com mais que 3 palavras.

Para medir os resultados, foram utilizadas as métricas: precisão, abrangência e f-measure. A precisão é sobre a capacidade da ferramenta de identificar os termos corretos, previstos, e a abrangência é sobre a quantidade de termos corretos extraídos. O método que apresentou o melhor balanço entre precisão e abrangência foi o de extração de sintagmas nominais – abordagem sintática –, com exclusão de termos por grupo semântico. A f-measure foi 11.51% para bigramas e 8,41% para trigramas. Além disso, de uma maneira geral, o método com exclusão de grupos semânticos obteve resultados melhores.

O número de termos extraídos pelo OntoLP é maior que o número de termos da referência, uma das explicações disso é que "os termos extraídos do *corpus* são considerados, sem a utilização de um ponto de corte por frequência" (Lopes et al., 2009, p. 82). Além disso, um número expressivo de termos da referência não foi detectado por nenhum dos métodos utilizados. Curiosamente, alguns termos relevantes extraídos pelo experimento não constaram na lista de referência. Isso pode ser explicado pelo fato de que em alguns casos "bigramas relevantes não foram constatados na referência, pois esta não inclui sub-termos. O bi-grama "aleitamento materno", por exemplo, está ausente da lista de referência, pois este é um sub-termo do o tri-grama "aleitamento materno exclusivo"" (Lopes et al., 2009, p. 83).

Outro trabalho a mencionar é o de Drouin (2003). Trata-se de mais um trabalho sobre extração sintagmas em *corpora* técnicos usando uma abordagem híbrida. O autor apresenta o TermoStat, ferramenta desenvolvida para a extração automática de termos que, por conta do uso da informação linguística, é capaz de restringir os itens lexicais que podem aparecer na lista.

Embora todos os trabalhos acima listados incluam aspectos linguísticos em seus métodos, tradicionalmente, os candidatos a termo de um *corpus* são extraídos de forma estatística, por meio de comparação entre um subcorpus e um *corpus* inteiro. A pesquisa de Drouin (2003) compartilha dessa metodologia comparativa. Sua ferramenta atua através da comparação de dois *corpora* distintos: um *corpus* de referência, que não é técnico (e sim mais geral), e um *corpus* de análise – composto pelos textos técnicos. A ideia é comparar a frequência de palavras em ambos os *corpora* e extrair apenas os termos específicos de uma área. No trabalho de Drouin (2003) o *corpus* de referência foi composto por 13.746 artigos de jornal, enquanto o *corpus* de análise, possuía três subcorpus diferentes voltados para o domínio de telecomunicações.

O que o autor intencionou foi saber quais são os itens que aparecem mais no *corpus* de análise do que seria o previsto, considerando o *corpus* geral. Para tanto, a pesquisa se baseia em uma técnica que dá acesso a dois critérios para quantificar a especificidade dos itens no *corpus*: valor de teste e probabilidade. A pesquisa seguiu o primeiro critério, que é uma visão padronizada da frequência das unidades lexicais. O autor utilizou o limite de "valor de teste" de +3,09, o que quer dizer que a chance de encontrar a frequência observada é menor que 1 em 1.000. Além da imposição deste limite, também foi imposta uma restrição linguística, através da anotação de POS: o algoritmo só pôde recuperar substantivos e adjetivos.

Os elementos extraídos, um subconjunto sobre o vocabulário específico do *corpus*, foram chamados de *pivôs lexicais especializados* (SLPs). A lista desses elementos foi revisada manualmente por três terminólogos profissionais, que avaliaram se o item era representativo do domínio ou do assunto do *corpus*. Para o autor, assim como para este trabalho, o resultado pode ser usado como peça central, e inicial, em um processo construção de uma terminologia.

Um parênteses é que o autor resolveu estudar o impacto do *corpus* de referência no resultado. Ele dividiu o RC em três *corpora* e repetiu o método para extrair novos SLPs. O resultado não mostrou uma variação significativa em termos

de frequência, mas mesmo assim significa que o *corpus* de referência usado tem alguma influência no resultado.

Dando sequência, o autor analisou os SLPs. É interessante que o tamanho das entradas lexicais também foi considerado: no máximo seis palavras. Isto pôde resultar na exclusão de alguns itens dos SLPs. De acordo com o trabalho, pesquisas mostram que quanto maior o tamanho de um item na lista, menor probabilidade de ser uma unidade terminológica válida. E, por curiosidade, sintagmas de duas palavras (palavras compostas) se mostram o tipo de termo mais frequente. A estrutura potencial dos termos candidatos (tendo em vista a língua inglesa) pode descrita dessa forma: (A|N)? (A|N)? (A|N)? (A|N)? (A|N)? N. Nessa estrutura, "A" corresponde a um adjetivo e "N" a um substantivo. De mais a mais, a lista final será filtrada, com o fim de encontrar termos fragmentados e de excluí-los.

Os resultados passaram por dois processos de validação, um automático e outro com interação humana. O processo automático consistiu em uma comparação entre termos candidatos obtidos e os termos da base de dados de uma terminologia. A leva que sobrou de termos candidatos foi submetida à validação de três especialistas da área de telecomunicações. A eles foram dados a lista de termos candidatos e o *corpus* de análise.

Quanto à abrangência, foi computada pelo número de termos candidatos extraídos pelo TermoStat sem usar a lista SLPs. O autor apresenta um resultado de 86.6%. A precisão geral obtida no conjunto de SLPs foi de 81%. Considerando os resultados positivos e a facilidade de uso do TermoStat, esta ferramenta foi utilizada, no escopo do presente trabalho, como uma das maneiras de obter termos do domínio do óleo e gás.

3.2. Extração automática de relações taxonômicas

A ideia desta seção é mencionar trabalhos que lidam com obtenção automática de relações taxonômicas, partindo tanto de *corpus*, como de terminologias. Nesse assunto, temos a obrigação de mencionar a pesquisadora Marti Hearst (1992), já que seu trabalho é um dos trabalhos mais emblemáticos no que se refere a uma abordagem linguisticamente motivada para a extração de

informação em *corpus*. Antes de apresentarmos propriamente o trabalho, contudo, convém esclarecermos algumas diferenças.

Nesta dissertação, objetivamos identificar e extrair, de um *corpus*, sintagmas ou palavras com potencial de pertencerem a uma terminologia voltada para óleo e gás. A partir das entradas geradas, pretendemos construir, de forma automática, uma taxonomia da área. Já o trabalho de Hearst (1992) segue um caminho diferente: apesar de que, como o nosso, ele também trata da extração de informação por meio de *corpus*, ele pretende extrair diretamente do *corpus* as relações de hiperonímia – taxonômicas – entre palavras.

Um dos objetivos da pesquisa de Marti Hearst (1992) era o de aumentar as relações existentes na WordNet – uma grande base lexical, construída manualmente, voltada para língua inglesa. Além disso, Hearst (1992) também visava auxiliar o trabalho de construtores de bases de conhecimento dependentes de domínio e lexicógrafos, que trabalhavam manualmente, sem o amparo de algo que automatizasse o processo. Para isso, a autora propôs a identificação de padrões léxico-sintáticos que contivessem uma relação buscada. A ideia era encontrar relações de hiperonímia por meio de padrões linguísticos, isto é, de forma automática.

A metodologia proposta por Hearst (1992) presume, então, a descoberta de tais padrões e a codificação de regras – por meio de anotação linguística – que deem conta de extraí-los. Isso significa que grande parte do trabalho é feita por meio de observação do *corpus*, o que pode tomar tempo. A autora sugere, para facilitar o trabalho, procedimentos padrões de descoberta, tais como: (1) ter clareza de qual é a relação buscada, no caso de Hearst, hiperonímia; (2) encontrar pares de palavras que contenham a relação de interesse; (3) extrair frases em que os pares apareçam e (4) tentar fazer generalizações com base nos contextos. Seguindo estes procedimentos em seu trabalho, os padrões propostos por Hearst, no que concerne a relação de hiperonímia foram:

(i) NP0 such as NP1 {, NP2 ... , (and | or) NPi}

(ii) such NP0 as {NP ,}* {(and | or)} NP

(iii) NP {, NP}* {,} or other NP0

(iv) NP {, NP}* {,} and other NP0

(v) NP0 {,} including { NP ,}* {or | and} NP

(vi) NP0 {,} especially { NP ,}* {or | and} NP

Nos padrões listados acima, NP0 significa sintagma nominal hiperônimo, enquanto os outros NPs significam os sintagmas nominais hipônimos. A detecção de tais padrões no *corpus* significa a extração de relações de hiperonímia, a princípio, corretas. No entanto, nem sempre, relações corretas são relevantes. No trabalho de Hearst, como resultado, 63% das relações extraídas foram consideradas possíveis de serem inseridas na WordNet. Parte dos resultados foi descartada, após uma análise comparativa entre as relações extraídas automaticamente e as relações contidas na WordNet.

Ao analisar seus resultados, a autora comenta sobre algumas dificuldades da tarefa. Percebeu-se que as relações de textos jornalísticos – gênero com o qual ela trabalhou –, tendem a ser menos prototípicas que as de textos enciclopédicos. Ou seja, textos de linguagem mais geral podem conter relações menos taxonômicas que terminologias. Além disso, a detecção automática das relações, apesar de boa precisão, tem seu ponto fraco na abrangência. As relações encontradas pelas regras costumam ser corretas, porém, se há relações relevantes fora delas, essas relações não são identificadas.

Mais recentemente, encontramos no SemEval diversas avaliações voltadas para extração de relações taxonômicas. Como o nome já denuncia, o SemEval (Semantic Evaluation) – tal como o CoNLL – consiste em uma série de workshops e avaliações anuais de sistemas, que, por sua vez, se voltam para a resolução de tarefas do interesse do PLN. O grande diferencial do SemEval, com relações às competições já descritas mais detalhadamente, é que ele é exclusivamente voltado para tarefas e questões semânticas da área.

No que se refere ao tema da seção, podemos mencionar duas edições que discutem relações taxonômicas. A primeira delas ocorreu em 2015 e foi a primeira sobre extração automática de taxonomias (Bordea et al., 2015). No SemEval-2015, a tarefa dos sistemas era a de, com base no material oferecido, encontrar relações de hiperonímia entre termos e estruturá-los hierarquicamente.

De acordo com Bordea et al. (2015), a elaboração de uma taxonomia segue três passos: extração dos termos, descoberta de relações entre os termos e construção da taxonomia. No SemEval-2015 (Bordea et al., 2015), os participantes puderam se concentrar no segundo e terceiro passo, uma vez que listas de termos foram oferecidas. A tarefa podia ser realizada em quatro diferentes domínios, dois de conhecimento geral – alimentação e equipamentos – e dois de domínios

especializados – química e ciência. Os participantes podiam submeter nos quatro domínios, se quisessem. Além disso, dois recursos dourados (gabaritos) foram produzidos para cada domínio: um com base na WordNet, e outro com base nas combinações de outros recursos linguísticos do domínio. Desse modo, os participantes também puderam submeter duas vezes em cada domínio.

O material de teste dos sistemas consistiu em oito listas de termos (terminologias) do domínio. Os participantes deveriam estruturá-las como uma taxonomia. Basicamente, eles deveriam retornar uma lista de pares com hiperônimo e hipônimo. Os sistemas foram avaliados com base na comparação dos resultados com dois *baselines* e com os gabaritos, recursos dourados, preparados pela organização. Diversas medidas foram utilizadas na validação, um dos métodos de avaliação consistia na comparação da estrutura gráfica dos resultados, isto é, os resultados foram comparados quanto à quantidade (tamanho) e qualidade de nodos e arestas. Esta avaliação verificava também se não havia sido produzido ciclos no resultado. De acordo com Bordea et al. (2015), uma propriedade das taxonomias é a ausência de ciclos, que, por sua vez, são inconsistentes com a ideia de relações hierárquicas.

Como as taxonomias-gabarito não eram completas, isto é – os sistemas eram capazes de identificar relações corretas e não previstas –, também foram realizadas avaliações manuais das relações novas propostas pelos sistemas. Em torno de 800 pares por sistema foram analisados. Para satisfazerem a correção, as relações deviam ser corretas e não podiam ser genéricas (Bordea et al., 2015).

Seis sistemas participaram da avaliação. Todos os sistemas usaram *corpus*, embora o SemEval não tenha oferecido, para descobrir relações. Algumas das abordagens mais comuns utilizadas pelos sistemas foram: (1) utilização de informação linguística, especialmente o método de detecção de padrões linguísticos, proposto por Hearst (1992); (2) estudo da co-ocorrência de informação em sentenças e documentos; (3) *substring matching;* e (4) extração de hiperonímia em definições presentes em recursos linguísticos.

Em 2016, foi realizada a segunda edição desta mesma tarefa – extração de taxonomia – no SemEval (Bordea et al. 2016). Da primeira edição para a segunda, poucas coisas mudaram. Quanto ao objetivo, ele permaneceu o mesmo: extrair relações hierárquicas de recursos linguísticos e estruturá-las hierarquicamente. Como podemos perceber, os participantes continuaram focando nos dois últimos

passos da extração de taxonomia, uma vez que listas de termos novamente foram disponibilizadas.

Um diferencial do SemEval-2016 (Bordea et al. 2016) é que ele disponibilizou *corpus* aos participantes, para que eles, se quisessem, extraíssem relações por essa fonte. Além disso, a segunda edição da tarefa deixou de ser restrita ao inglês e se estendeu para mais outras três línguas: francês, italiano e alemão. De mais a mais, os domínios em questão também se modificaram e diminuíram para três: meio-ambiente, alimentação e ciência. Com relação ao material de teste, ele consistiu em seis listas para cada domínio específico e para cada língua.

Agora, no que concerne a preparação de gabaritos, foram produzidos recursos dourados para os três domínios em questão. Isso foi feito a partir da Wikipédia, da Eurovoc, da WordNet e a partir de outros recursos de domínio específicos. O grande desafio foi a produção de gabarito para as outras três línguas. Basicamente, os recursos dourados do inglês, de cada domínio, foram traduzidos manualmente por seis linguistas. Esse processo não foi tão simples, já que nem todos os conceitos são traduzíveis de uma língua para outra e, além disso, as relações entre os conceitos, com a transposição da língua, também eram variáveis.

Cinco times participaram da competição, o que implicou em 62 sistemas submetidos. É válido lembrar que os participantes podiam submeter em todas as possibilidades de domínio, língua e tipos de gabarito. Quanto à avaliação dos sistemas, ela levou em conta os mesmos métodos da edição de 2015: comparações com os recursos dourados – utilizando medidas de precisão e abrangência –; análise estrutural qualitativa e quantitativa dos resultados, analisando o tamanho das taxonomias geradas, bem como os nodos e arestas; e uma avaliação manual para analisar novas relações não previstas – foram validadas 6.200 pares de relações novas no total.

Todos os sistemas passaram por essas três validações, no entanto, o rank final foi decidido com base em sete propriedades que abrangiam todas as avaliações: (1) presença/ausência de ciclicidade; (2) similaridade estrutural (tamanho) com o gabarito; (3) categorização (categorias taxonômicas); (4) conectividade – sobre a quantidade de componentes conectados ou soltos –; (5) sobreposição de arestas com a taxonomia-gabarito; (6) número de domínios cobertos pelo sistema e (7) precisão das novas relações validadas manualmente (Bordea et al. 2016). O sistema que se saiu melhor em quase todas as avaliações e

propriedades foi o TAXI (*TAXonomy Induction system*), que faz uso de uma abordagem baseada nos padrões Hearst (1992). De mais a mais, o fato de que nenhuma das taxonomias submetidas por ele apresentou ciclo contribuiu bastante para o resultado (Bordea et al., 2015).

3.3. GENIA E CRAFT

O objetivo desta seção é apresentar brevemente *corpora* em domínios especializados, anotados com entidade, por meio de recursos linguísticos. O primeiro a ser mencionado é o GENIA (Thompson et al., 2017), *corpus* do domínio biológico, constituído por 2 mil resumos da MEDLINE - base de dados da área médica -, 400 mil palavras e quase cem mil anotações de termos biológicos. Como já foi mencionado, o GENIA foi criado devido à ausência de *corpora* extensivamente anotados na área, com a finalidade de ser referência para as técnicas de PLN.

O *corpus* GENIA possui multicamadas de anotação e foi anotado manualmente. Ao longo dos anos, as anotações do *corpus* foram continuamente enriquecidas, fazendo com ele se tornasse um *corpus* amplamente utilizado no treinamento de vários sistemas do domínio biomédico. Algumas das camadas de anotação do GENIA são: entidades, POS, correferência, análise sintática, eventos, relações, e anotação de "meta-conhecimento" (Meta-Knowledge Annotation) – que funcionaria como uma interpretação dos eventos ou como informações relacionadas ao discurso (Thompson et al., 2017).

No contexto desta pesquisa, o que mais interessa é a anotação das entidades, isto é, a anotação dos termos relevantes do domínio biomédico, que, por sua vez, foi feita com base em uma ontologia criada manualmente partir do *corpus*, pelos próprios organizadores. Antes de mais nada, foi feita uma busca por recursos semânticos existentes na área, tais como taxonomias e ontologias. Entretanto, os recursos encontrados acabaram por apresentar diversas questões e desafios, o que levou aos organizadores a criar uma nova e própria ontologia – a partir do *corpus* – que conteria, de forma hierarquicamente organizada, os termos do GENIA (Thompson et al., 2017).

Em outras palavras, os organizadores elaboraram um *corpus* e uma ontologia ao mesmo tempo, sendo a última utilizada como recurso de anotação do *corpus*. A ontologia criada possui 47 categorias nominais relevantes na biologia e diversos termos considerados conceitos biológicos. Algo interessante é que todos os termos foram classificados a partir de apenas duas categorias: *substance* e *source*. Os termos se originam de hierarquias que se sucedem, por exemplo: *source* > *natural* > organism > virus. Por outro lado, temos: *substance* > *compound* > *organic* > *carbohydrate*. O número total de termos recuperados, isto é, que integram a ontologia, é 93.293. Distribuindo o número de termos recuperados, temos: 89.862 termos simples e 3.431 termos compostos (KIM et al., 2003).

Uma particularidade do domínio biológico destacada no texto, no que concerne à anotação dos termos relevantes, é o fato de que, na biologia, as entidades mencionadas são substantivos comuns (por exemplo: vírus), que aparecem com uma variedade de especificadores e qualificadores, enquanto que entidades mencionadas de textos gerais normalmente são substantivos próprios sem muito modificadores. Por isso, no GENIA, os especificadores não foram incluídos nos termos e a inclusão de qualificadores foi deixada para avaliação dos especialistas. Isso talvez explique a razão de haver uma quantidade muito maior de termos simples do que compostos no *corpus* (KIM et al., 2003).

Já um desafio da tarefa, que foi mencionado, tem a ver com os termos que aparecem em estrutura de coordenação, que, na verdade, possuem dois termos, mas um com elipse. Por exemplo, em “receptores CD2 e CD5”, teríamos, em teoria, os termos “receptor CD2” e “receptor CD5”. Tal dificuldade também apareceu no trabalho de Wendt et al. (2010), com entidades da geologia. Em Wendt et al. (2010), os autores optaram por eliminar a segunda parte do termo, que está coordenada. Já no GENIA, pela ontologia ter sido criada manualmente por especialistas, a partir de *corpus*, os casos de coordenação foram considerados termos separados (KIM et al., 2003).

O GENIA é um trabalho de referência no contexto desta pesquisa por duas especiais razões: (1) é um *corpus* que partiu da construção de um recurso lexical, ontologia, para fazer a anotação das entidades, e (2) é um *corpus*, de domínio técnico, que supriu a ausência de material de referência na biomedicina para técnicas do PLN e se tornou amplamente utilizado no treinamento de sistemas. De forma parecida, o presente trabalho lida com dois objetivos: um mais palpável, (1)

auxiliar a identificar e classificar as entidades de um *corpus* em uma área técnica – elaborando uma taxonomia na área de óleo e gás que auxilie nas tarefas –, e outro mais geral, (2) prover material de referência para o PLN, no domínio de O&G, em língua portuguesa.

Outro *corpus* a ser mencionado nesta seção é o CRAFT (Cohen et al., 2017), um *corpus* composto por artigos de biomedicina e com diversas camadas de anotação. O CRAFT possui as mesmas motivações do GENIA e é inspirado por ele. No entanto, ele surgiu pela necessidade de se trabalhar com artigos completos (outros gêneros) e não apenas com resumos, como é o GENIA. Os organizadores perceberam que no corpo do texto havia mais informações e que a estrutura textual era diferente, isto é, os resumos e as outras seções de textos acadêmicos possuem características diferentes. Além disso, a maioria dos trabalhos de PLN, na área de biomedicina, focalizavam em apenas duas categorias de entidade mencionada, seguindo o padrão do GENIA. No CRAFT, havia o interesse em ampliar as possibilidades de classes semânticas.

Assim como o GENIA, o CRAFT possui diversos tipos de anotação: anotação morfossintática, entidades mencionadas e correferência. O processo de anotação foi diferente para cada camada, mas as anotações, em geral, contaram com o apoio de estudantes de graduação e pós-graduação, tanto por especialistas da área, como por linguistas.

Sobre as anotações morfossintáticas, as categorias lexicais foram feitas pelo anotador GENIA e a estrutura sintática foi etiquetada pelo anotador OpenNLP. Posteriormente, a anotação foi revista e corrigida por anotadores. Quanto à anotação de entidades mencionadas, ela foi feita com base em sete ontologias, que continham mais de cem mil conceitos. A princípio, foi utilizado um método de correspondência entre as palavras, mas a anotação também teve participação humana e foi feita uma concordância entre anotadores (Cohen et al., 2017).

Já a correferência, foi anotada por estudantes de linguística e biologia, sem apoio de informação sintática. A anotação de correferência se inspirou nas diretivas do OntoNotes (HOVY et al., 2006), um *corpus* de amostra geral da língua, com várias camadas de anotação, variados gêneros textuais e diversas línguas. O OntoNotes é um *corpus* frequentemente tomado como referência para treino de sistemas. Neste ponto, é interessante destacar que Cohen et al. (2017) dá ênfase ao

fato de as técnicas de PLN de *corpora* gerais não funcionarem bem em domínios especializados.

No caso da correferência, os autores evidenciaram que a resolução automática é mais difícil no domínio biomédico e que os sistemas precisam de uma alteração significativa para funcionarem bem. Segundo os autores, um dos possíveis fatores da dificuldade seja o fato de os artigos biomédicos terem tamanhos muito maiores que textos jornalísticos em geral. Consequentemente, os artigos têm cadeias de correferência maiores. No CRAFT, os anotadores trabalharam por parágrafo e depois ligaram as entidades do parágrafo a menções anteriores no documento (Cohen et al., 2017).

Uma das diferenças entre as diretivas de anotação do OntoNotes e do CRAFT é que 'termos genéricos', no OntoNotes, não são marcados. Já no CRAFT, estes termos não existem, pois considerar um termo como genérico ou abstrato pode ser prejudicial para a correferência do domínio biomédico. Por isso, todos os nomes no CRAFT foram tratados como marcáveis. Exemplo: "Allergan Inc. said it received approval to sell the Phaco Flex intraocular lens, the first foldable silicone lens available for cataract surgery. The lens' foldability enables it to be inserted in smaller incisions than are now possible for cataract surgery". Nesta frase, não foi marcada correferência em "cataract surgery" pelo OntoNotes, mas teria sido pelo CRAFT.

Tanto o GENIA e o CRAFT foram mencionados aqui por serem bastante representativos para suas áreas. Ambos são *corpora* referenciados no domínio biomédico, o primeiro, inclusive, é um projeto que começou no início do século e que continua sendo aprimorado. Além de também serem voltados para o domínio técnico, os projetos se relacionam com a presente pesquisa porque partem de recursos lexicais para realizarem as anotações do PLN. Do mesmo modo, a taxonomia proposta aqui se propõe a ser um ponto de partida para anotações em um *corpus* de óleo e gás.

4
Metodologia da pesquisa

Finalmente, este capítulo tem como objetivo descrever todo o processo realizado para a construção de uma taxonomia na área do óleo e gás. Como já foi ligeiramente mencionado, o processo contou com duas grandes etapas, sendo que, na primeira, fizemos uma listagem de termos relevantes do domínio – ou seja, elaboramos uma possível terminologia da área –, e, na segunda, agrupamos os itens da lista por relações taxonômicas (de 'mãe e filho'), disponibilizando uma organização e visualização hierárquica e estruturada da terminologia. O intuito da criação de uma taxonomia é oferecer um recurso linguístico que possa ser usado como insumo em tarefas do PLN que envolvam anotações linguísticas, sobretudo de entidades.

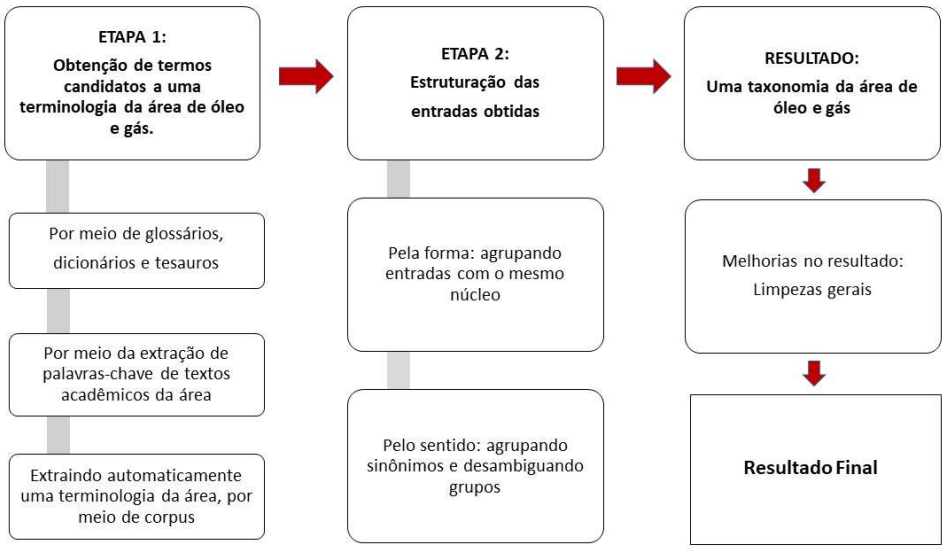

Figura 1 – Metodologia da pesquisa

4.1.
Obtenção de candidatos a termos no domínio

O primeiro passo para a identificação de entidades mencionadas em *corpus* de domínio técnico é a listagem de termos candidatos. Entretanto, listar termos

candidatos a entidade implica, de algum modo, listar termos candidatos também a uma terminologia da área. Para se elaborar uma terminologia, contudo, um primeiro passo é o reconhecimento do domínio em questão, no nosso caso: o domínio de óleo e gás. Na figura abaixo, apresentamos uma generalização de classes relevantes do domínio de óleo e gás.

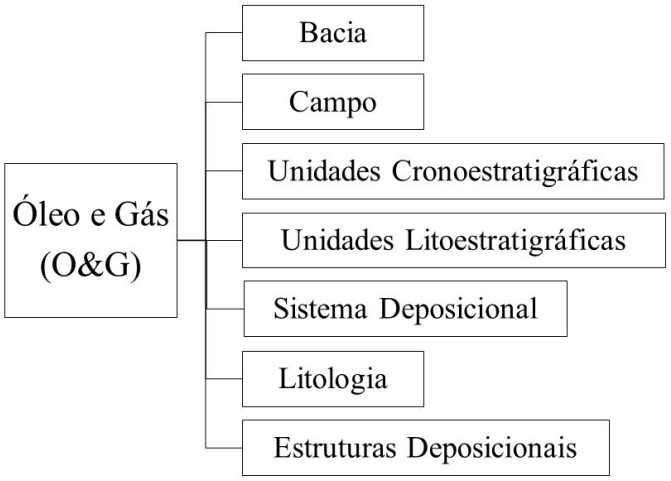

Figura 2 – Árvore de domínio

O esquema acima foi construído com o apoio de especialistas da área e expressa uma determinada visão sobre o domínio em foco. O quadro serviu não para determinar os termos candidatos ou estrutura da taxonomia, mas como uma ideia inicial de uma possível organização do domínio.

A partir disso, inicialmente, buscamos reunir o vocabulário do domínio em xeque. Elaboramos uma lista com candidatos a termo, partindo de três fontes distintas: (1) vocabulários especializados, (2) um *corpus* específico do domínio, e (3) metadados do *corpus*: palavras-chave de teses, dissertações e monografias.

Os vocabulários, por serem repertórios terminológicos feitos por especialistas, são úteis para se ter acesso, com confiança, aos termos relevantes ou relacionados ao domínio. Por isso, neste trabalho, todas as entradas dessas fontes foram consideradas candidatas a entidades e entradas na taxonomia. Utilizamos cinco vocabulários, de diversos tipos, disponíveis na internet: (1) Tesauro da Indústria do Petróleo Volume 1[1], (2) Tesauro da Indústria do Petróleo Volume 2[2],

[1] https://silo.tips/download/tesauro-da-industria-do-petroleo-2#
[2] https://silo.tips/download/tesauro-da-industria-do-petroleo

(3) Glossário da ANP[3], (4) Dicionário de Petróleo da Língua Portuguesa[4], e (5) Glossário da Indústria do Petróleo e Gás[5].

Já o *corpus* em si, o Petrolês, - que será explicado mais adiante - também foi utilizado para descobrir o que poderia contar como entidade no domínio. Nesse caso, buscamos extrair os termos relevantes a partir de duas ferramentas: AntConc e TermoStat – ambas já mencionadas no capítulo anterior. A primeira possui uma abordagem totalmente estatística, enquanto a segunda utiliza um método híbrido, com abordagens linguística e estatística. Para a identificação de termos relevantes e palavras-chave, ambas funcionam de forma parecida: um *corpus* de interesse (no nosso caso, de domínio específico) é comparado com um segundo *corpus* (*corpus* de contraste). As palavras de um e outro são confrontadas, em termos de frequência, e, assim, os termos relevantes do *corpus* de interesse são extraídos.

Quanto às palavras-chave, elas se referem àquelas utilizadas pelos autores dos textos acadêmicos que compõem o *corpus* Petrolês. ou seja, são as palavras-chave de em torno de 400 teses e dissertações de óleo e gás. Como essa lista também foi criada partindo de sintagmas escolhidos por pessoas especialistas na área (pesquisadores e autores de publicações acadêmicas), todos os itens, a princípio, foram considerados.

É importante frisar que a extração de palavras-chave de que nos referimos aqui não é aquela utilizada para se extrair automaticamente *keywords* de grandes *corpora* – como é feito, por exemplo, pela ferramenta AntConc, que será explicada mais adiante. Estamos falando das palavras-chave dos textos/artigos acadêmicos. A extração foi feita com base nos metadados de um *corpus* composto por teses, dissertações e monografias, não no *corpus* em si.

A seguir, descrevemos mais detalhadamente os procedimentos realizados com base em cada uma das fontes, mas antes de avançarmos é importante lembrar que as listas das duas primeiras fontes – vocabulários e palavras-chave –, por se tratarem de itens afirmados por especialistas, foram considerados de alta confiança. Já os itens da terceira fonte, por terem relação com um processamento automático - que não foi feito nem validado por pessoas da área –, foram considerados de confiança mais baixa. Além disso, adicionamos ao resultado – que é a união destas

[3] https://www.gov.br/anp/pt-br/acesso-a-informacao/glossario
[4] http://dicionariodopetroleo.com.br/
[5] https://www.presalpetroleo.gov.br/ppsa/legislacao/glossario-da-industria-de-petroleo-e-gas

três fontes – um siglário[6] da área de óleo e gás. Apesar de ter não ter contado como um glossário, o siglário foi tomado como sendo de mesmo valor, uma vez que é uma lista pronta feita por especialistas e que pode conter entradas presentes no *corpus*.

4.1.1. Vocabulários

Criamos uma lista composta pelo somatório das entradas (apenas verbetes, sem definições) de cinco vocabulários públicos no domínio: Tesauro da Indústria do Petróleo Volume 1, Tesauro da Indústria do Petróleo Volume 2, Glossário ANP, Glossário da Indústria do Petróleo e Gás e Dicionário do Petróleo em Língua Portuguesa.

Os dois tesauros utilizados foram organizados pelo Centro de Pesquisas e Desenvolvimento Apoio à Gestão/ Informação (Cenpes), vinculado à Petrobrás. Ambos são bilíngues e não possuem definições. Entretanto, diferenciam-se quanto aos temas abordados. O volume 1 é voltado para transporte, refino e petroquímica, enquanto o volume 2 é voltado para exploração, perfuração e produção.

O glossário ANP, na verdade, não tem um nome, mas o denominamos assim porque ele é organizado pela Agência Nacional do Petróleo, Gás natural e Biocombustíveis (ANP). Trata-se de um glossário monolíngue e com definições. Já o glossário da Indústria do Petróleo foi elaborado no âmbito do site do Pré-sal. É monolíngue e com definições. No enquadre desta pesquisa, as definições dos glossários não foram utilizadas.

Por último, o Dicionário do Petróleo em Língua Portuguesa, apesar do nome, não apresenta definições, apenas os verbetes. No entanto, é uma nomenclatura bilíngue (inglês) e traz variações da língua portuguesa do Brasil, Portugal e Angola. É voltado para exploração e produção de petróleo e gás e possui diversos editores: Eloi Fernández y Fernández, Oswaldo A. Pedrosa Junior, Antônio Correa de Pinho.

Os cinco documentos foram unificados como uma lista só, e, em seguida, procedimentos semiautomáticos foram realizados com o fim de padronizar as entradas. Antes dos procedimentos, as nomenclaturas possuíam 22.029 candidatas

[6] http://dicionariodopetroleo.com.br/siglario

a entidades mencionadas. Após o tratamento, obtivemos 19.079 entradas. Essa é a fonte que gerou o maior número de candidatas. Abaixo listamos os procedimentos semiautomáticos realizados com a lista dos glossários:

- As entradas foram uniformizadas com letra minúscula e colocadas em ordem alfabética, dado que, em um primeiro momento, eliminar a diferenciação entre letras minúscula e maiúscula foi necessário para descartar entradas repetidas.
- Tudo o que veio depois da entrada, e que também não era a definição, foi eliminado – informações entre parênteses, colchetes, chaves travessões e vírgulas. Exemplos:

indício de gás e indícios de gás ~~(port.)~~
gás natural liquefeito ~~(gnl)~~
indexação ~~(perfuração)~~
anel ~~[dispositivo de acoplamento para unir dois tubos]~~
loran ~~{determinação de posição de navio no mar a longa distância}~~
biodiesel ~~– b100~~
entregue no terminal ~~– dat~~
gás natural comprimido ~~– gnc~~
área de proteção ambiental~~, brasil,~~
instituto nacional da propriedade industrial ~~(inpi), brasil,~~
lei do petróleo~~, brasil~~

- Itens repetidos foram excluídos.

Além disso:

- Entradas com preposições diferentes não foram unificadas. Exemplo:

Jateamento com areia
Jateamento de areia

- Diferenças de plural e singular não foram padronizadas.
- Entradas com ‘ou’, ‘e’ e ‘/’ foram pesquisadas manualmente e mantidas, com a intenção de serem tratadas em um momento posterior. Além disso, alguns casos especiais de vírgula também foram mantidos (casos de coordenação). Exemplos:

agência nacional do petróleo, gás natural e biocombustíveis
ampliação (ou expansão ou aumento) da capacidade de transporte

análise de modos, efeitos e criticidade de falhas
assistência a pré-operação, partida e operação
desmontagem, movimentação e montagem
desmontagem, transporte e montagem
urânio e minerais aliados
contato gás/água
custos do pis/cofins
dados de tempo/profundidade

4.1.2. *Corpus*: Petrolês

O corpus foi criado a partir de 409 documentos, dentre monografias, teses e dissertações, da área de óleo e gás. Especificamente, trata-se de um subconjunto do material descrito em Gomes et al. (2018), que contém documentos públicos disponibilizados por instituições de referência na área, especificamente Petrobrás e Agência Nacional do Petróleo, Gás e Biocombustíveis (ANP). O corpus contém cerca de 6 milhões de tokens, nenhuma anotação e poucos cuidados no pré-processamento.

4.1.3. *Corpus*: extração de sintagmas relevantes

Para a extração automática, por meio de *corpus*, de elementos candidatos a termos da terminologia, partimos de duas ferramentas distintas: AntConc e TermoStat. Enquanto a primeira possui uma abordagem totalmente estatística, a segunda utiliza um método híbrido, apoiado em informação linguística.

O TermoStat (Drouin, 2003), ferramenta disponível na web e cujo funcionamento já foi brevemente ilustrado na seção 3.1, faz uma análise comparativa entre dois *corpora* distintos. Recapitulando o seu funcionamento, trata-se de uma comparação de frequência (que leva em conta anotação de POS) de palavras entre um *corpus* de referência (de linguagem geral e que já vem embutido com a ferramenta) e um *corpus* de análise (*corpus* de domínio). O *corpus* de referência da ferramenta é formado por textos jornalísticos – e não técnicos – de Portugal.

A comparação de frequência é feita com um limite de valor de teste de +3,09, para extrair só o que é específico do *corpus* de análise, isto é, para extrair os sintagmas que são específicos de um domínio. Entretanto, nessa comparação, também é imposta uma restrição gramatical, que permite ao usuário escolher quais classes de palavras a ferramenta deve extrair. Dessa forma, a anotação linguística de POS e a comparação de frequência andam juntas, como ponto de partida do processo.

Então, através da análise comparativa, a ferramenta gera uma lista de palavras com as classes escolhidas, tais palavras extraídas são consideradas *headwords*, e é a partir delas que a ferramenta busca por sintagmas. É como se ela fosse buscar no *corpus* por sintagmas onde *headwords* estão. Durante a identificação dos sintagmas, o TermoStat limita o tamanho do sintagma que irá considerar, impondo no máximo seis palavras.

No caso da presente pesquisa, o TermoStat fez uma análise comparativa entre um *corpus* de contraste e o *corpus* de 409 documentos acadêmicos de óleo e gás, que carregamos no programa. De maneira geral, a análise pode ser feita – a escolha é do usuário – por meio de diferentes cálculos estatísticos, entre eles: frequência bruta, especificidade, teste X2, Log-likelihood e Log-odds ratio. No nosso caso, optamos pelo critério de "Especificidade", criado por Lafon (1980) e proposto para definir um vocabulário específico de um subcorpus comparado ao conjunto de um *corpus*.

Como resultado, obtivemos uma lista com 9.033 candidatas a termos (ou a EM). Em termos comparativos, a lista gerada pelo GlossEx, de Park et al. (2002), possui 9.862 itens, sem incluir os itens variantes, mas o *corpus* utilizado possui em torno de 200 mil palavras. Nosso *corpus*, de 400 documentos, tem mais de 6 milhões de tokens. Sem validação ou revisão, a lista gerada pelo OntoLP, do trabalho de Lopes et al. (2009), gerou 3.645 itens, sendo que trabalhou com um *corpus* de 283 textos, com 785.448 palavras. No resultado final do trabalho, o número caiu para 2.407 itens. Já o trabalho de Wendt et al. (2010), com o ExATOlp, gerou uma lista de 4.626 itens, partindo de um *corpus* de 140 textos com aproximadamente um milhão de palavras. Abaixo, uma tabela comparativa do resultado das ferramentas.

TAMANHO DO CORPUS	FERRAMENTA UTILIZADA	QUANTIDADE DE TERMOS CANDIDATOS IDENTIFICADOS
225.100 mil palavras	GlossEx	9.862 (Park et al., 2002)
785.448 palavras	OntoLP	3.645 (Lopes et al., 2009)
1 milhão de palavras	ExATOLp	4.626 (Wendt et al., 2010)
6 milhões de tokens	TermoStat	9.033 (este trabalho)

Quadro 1 – Ferramentas de extração de termos

É válido mencionar que a comparação é um pouco injusta, já que cada trabalho lidou com uma situação específica e atuou com uma metodologia própria. Diferentemente dos outros trabalhos, os resultados do TermoStat não sofreram nenhuma modificação ou tratamento. Nós extraímos os sintagmas nominais compostos do *corpus*, mas sem criarmos filtros, tais como regras de transformação, eliminação de certas classes gramaticais, artigo no início da entrada ou número restrito de tokens por entrada. As medidas de corte, que serão descritas mais adiante, foram tomadas apenas posteriormente, quando os vocabulários de todas as fontes foram unidos. Ainda assim, organizamos a tabela acima com o intuito de acompanhar o processo, de um ponto de vista quantitativo.

Corpus >> **verso_petrocorpus_1.0__unificada** wograine | Help | Log out

Results

List of terms | Cloud | Stat | Structuration | Bigrams

Candidate (grouping variant)	Frequency	Score (Specificity)	Variants	Pattern
óleo diesel	1246	33.26	*óleo diesel* *óleos diesel*	Common_Noun Common_Noun
gás natural	1418	32.6	*gás natural* *gases naturais*	Common_Noun Adjective
resultado obtido	1280	32.12	*resultado obtidos* *resultados obtidos*	Common_Noun Adjective
ácido graxo	1143	31.89	*ácido graxo* *ácido graxos* *ácidos graxos*	Common_Noun Common_Noun
rio grande	1190	31.87	*rio grande*	Common_Noun Common_Noun
t t	955	29.14	*t t*	Common_Noun Common_Noun
corpo de prova	821	27.02	*corpo de prova* *corpos de prova* *corpos de provas*	Common_Noun Preposition Common_Noun
fase líquido	820	27	*fase líquido* *fases líquido* *fases líquida* *fases líquidas*	Common_Noun Adjective
meio poroso	654	24.11	*meio poroso* *meio porosos* *meios porosos*	Common_Noun Adjective
óleo vegetal	635	23.58	*óleo vegetal* *óleos vegetais*	Common_Noun Adjective
produção de biodiesel	624	23.55	*produção de biodiesel* *produção do biodiesel* *produções de biodiesel*	Common_Noun Preposition Common_Noun
dado experimental	600	23.04	*dado experimental* *dados experimentais*	Common_Noun Adjective
desvio padrão	592	22.93	*desvio padrão* *desvios padrão*	Common_Noun Common_Noun

Figura 3 – TermoStat

Cabe mencionar que, além de possibilitar variedade de cálculos estatísticos, a ferramenta TermoStat apresenta os resultados de formas diversas, o programa disponibiliza: (1) a lista dos termos relevantes - que vêm com frequência,

especificidade, suas variantes (plural e singular) e padrão gramatical/linguístico; (2) nuvem de palavras mais frequentes; (3) os padrões linguísticos obtidos e a frequência de cada um deles e (4) exemplos de vezes em que os termos da lista aparecem em sintagmas maiores. Exemplos: gás natural > processamento de gás natural.

Já a funcionalidade *keyword* da ferramenta AntConc funciona de forma parecida com o TermoStat, porém com uma abordagem totalmente estatística e sem nenhum apoio de informação linguística. Durante a utilização do AntConc, o *corpus* de domínio específico utilizado foi o mesmo que no TermoStat. No entanto, o *corpus* de contraste não é dado pela ferramenta. Por conta disso, carregamos um *corpus* composto por textos jornalísticos de diversas áreas, incluindo arte, economia, política, ciência, educação, celebridades, etc.

Assim como no TermoStat, no AntConc é possível realizar a mesma tarefa com diferentes fórmulas: Log-likelihood (4 term) - a padrão do AntConc -; Log-likelihood (2 term); Chi-Squared (4 term) + Yates, Chi-Squared (2 term), Chi-Squared (2 term) + Yates e Chi-Squared (4 term). Testamos o experimento com algumas das fórmulas e no final optamos pela estatística Log-likelihood (4 term). Essa escolha não teve uma motivação forte, observamos que as duas primeiras fórmulas não resultaram em listas com expressiva diferença, ao contrário, geraram listas parecidas e de tamanhos iguais: 10.293 entradas. A fórmula Chi-Squared (4 term) + Yates obteve 8.531 entradas, sendo a menor lista, e Chi-Squared (2 term) obteve 8.924 entradas. No final, acabamos optando pela fórmula dada como padrão da ferramenta e a que obteve mais candidatas.

Os resultados do AntConc não foram considerados tão bons ou tão complexos quanto os do TermoStat. Uma das razões é que, diferentemente do TermoStat, o AntConc não extrai sintagmas compostos, todas as palavras-chave listadas são palavras simples (unigramas), o que, a princípio, parece uma desvantagem, levando em conta que em terminologias os termos compostos são muito relevantes (KRIEGER & FINATTO, 2004). Além disso, nos trabalhos mencionados como referência, são os sintagmas compostos que recebem mais atenção. Lopes et al., para construir uma ontologia, nem sequer trabalharam com unigramas. O trabalho de Wendt et al. (2010), com o ExATOlp, e o trablho de Park et al. (2002), com o GlossEx, geraram mais bigramas que unigramas.

Além disso, o TermoStat, como já dito, faz uso de anotação linguística de pos durante o processo – ainda que, para a língua portuguesa, essa funcionalidade apresente alguns problemas. O conhecimento linguístico permite não só oferecer maiores informações sobre os sintagmas extraídos – como o padrão linguístico, por exemplo – como ajuda o programa a considerar sintagmas com palavras mais significativas: substantivos, por exemplo. Por conta desta diferença, o AntConc, que não utiliza informação linguística, listou, por exemplo, com base em frequência, palavras de diversas classes gramaticais. Devido a essas razões, os resultados do AntConc se mostraram piores, e, por conta disso, optamos por não utilizar, em um primeiro momento, a lista gerada por ele.

No entanto, instigados pelo desejo de reaproveitar os dados do AntConc, decidimos também por refazer o procedimento no TermoStat, gerando desta vez apenas termos simples, unigramas. O intuito seria o de comparar as duas ferramentas, no quesito termos simples. Então, geramos uma lista do TermoStat com termos simples de variadas classes gramaticais: substantivos, adjetivos e verbos. A ideia era gerar uma lista que fosse comparável com a do AntConc, que, por sua vez, não limita classes gramaticais. A lista nova gerada trouxe 8.671 entradas, enquanto a do AntConc continha 10.293 itens. Comparando as duas listas, medimos 6.020 unigramas compartilhados. Isso significa que, unindo os resultados, 63% é informação de mesmo valor, e 36% constitui informação particular de cada ferramenta.

A princípio, esse resultado parece mostrar que as ferramentas, apesar de funcionarem de modos distintos, resultam em dados similares para a identificação de termos simples de um domínio. Isto é, pelo menos quando se consideram variadas classes gramaticais. A lista dos unigramas compartilhados entre TermoStat e AntConc não resultou em unigramas para a taxonomia, mas foi preservada, com a finalidade de possibilitar um experimento comparativo.

4.1.4. Metadados do *corpus*: palavras-chave

Criamos uma lista composta pelas palavras-chave de 409 textos acadêmicos – teses, dissertações e monografias – do domínio de petróleo. Ela foi extraída de forma automática dos metadados do *corpus* Petrolês, que contém seis milhões de tokens. A princípio, as palavras-chave do *corpus* geraram 1.444 possíveis entidades

mencionadas. Com o fim de eliminar palavras-chave repetidas nos documentos, a lista também passou por tratamento semi-automático e, ao final, obtivemos 1.131 candidatas. Abaixo listamos os procedimentos realizados:

- As palavras do tipo "tese", "monografia", "mestrado" e/ou "dissertação" foram retiradas quando foram usadas como palavra-chave ou quando apareciam ao lado de uma palavra-chave.
- "Ponto e vírgula" ao final das entradas foi excluído.
- Entradas foram padronizadas em minúscula.
- Erros de português ou de digitação foram corrigidos manualmente.
- Palavras com diferença de número ficaram no singular. Como a lista de palavras-chave é bem menor que a lista de glossário, a identificação e mudança pode ser feita manualmente, sem auxílio de anotação, o que não ocorreu nos glossários.
- Tudo o que veio depois da entrada foi eliminado, por exemplo, informações entre parênteses (como no caso dos glossários).
- Itens repetidos foram excluídos.
- Assim como nos glossários, entradas com preposições foram mantidas sem alteração e, além disso, não foram traduzidas palavras-chave de língua estrangeira.

4.1.5. Siglário

Adicionamos um siglário do domínio de óleo e gás, disponível na web, às nossas fontes. As siglas da área, contidas no siglário, foram igualmente consideradas como possíveis candidatas a entidade mencionada – junto com os glossários, palavras-chave e sintagmas extraídos do *corpus*. O siglário também passou por alguns procedimentos semiautomáticos, são eles: (1) as definições ou significados da sigla foram eliminados em um primeiro momento; (2) siglas com variações de minúscula para maiúscula foram apagadas, mantendo apenas uma variante (em maiúscula) e (3) siglas repetidas foram excluídas.

4.1.6.
Comparação entre as fontes

Com o que foi feito até agora, já é possível fazer um muito breve comentário avaliativo sobre o processo de gerar candidatas a entidades, partindo de diferentes fontes e métodos. Ao total, o somatório de todas as candidatas, oriundas de distintos meios, chegou a 30.326 entradas. Eliminamos as entradas duplicadas e ficamos com um número de 29.035 candidatas. Com isso, construímos uma primeira terminologia "leve" da área. Em comparação com a quantidade de termos obtidos em outros trabalhos (tabela acima), quase 30 mil entradas é muito. Porém, quando lembramos que a ontologia utilizada para a anotação do *corpus* GENIA contém 93.293 entradas (seção 3.2), parece, pelo contrário, que nosso método é econômico.

Algo que chamou a atenção foi que, depois de todos os procedimentos semiautomáticos realizados em cada uma das fontes – com exceção do TermoStat – voltados para eliminação entradas repetidas, apenas um pouco mais de mil entradas foram duplicadas ao se juntarem. Ou seja, nos glossários, nas palavras-chave e no TermoStat, havia 1.291 itens compartilhados e 29.035 itens únicos, o que significa que 4,2% de todos os valores eram compartilhados entre as fontes e 95,7% era informação singular. A tabela abaixo apresenta a contribuição de cada recurso ou método para a elaboração da lista de termos de domínio.

	QUANTIDADE
Candidatos pré-tratamento semiautomático	30.326
Candidatos pós-tratamento semiautomático	29.035
Itens compartilhados entre as fontes	1.291 (4,2%)
Itens únicos entre as fontes	29.035 (95,7%)

Quadro 2 – Candidatos a termos pelas diversas fontes

4.2.
Construção de uma taxonomia de óleo e gás

Após os procedimentos acima descritos, as listas, oriundas de diversas fontes, foram unificadas como uma lista única de itens candidatos a termos da taxonomia. Como foi colocado na última seção, por meio de comandos simples de editores de texto, os termos repetidos foram eliminados, gerando uma terminologia de 29.035 termos candidatos. Agora, nesta seção, detalhamos a segunda etapa do

processo de construção de uma taxonomia de óleo e gás. Tal etapa consiste em criar uma organização para o repertório terminológico obtido, fazendo com que os itens, que se encontram apenas listados, numa estrutura linear, possam ser organizados em relações de hiperonímia, o que é a característica mais predominante de uma taxonomia.

Assim, com o auxílio da programação[*], uma nova visualização das entradas foi elaborada visando compreender a estrutura subjacente à lista, ou seja, buscamos uma visualização hierárquica dos itens contidos, que nos permitisse entender melhor o que havíamos conseguido com 29.035 mil candidatos. Inicialmente, organizamos nossa lista pelo critério formal, agrupando sintagmas com mesmo núcleo, repetindo o processo de Freitas (2007). Em seguida, ampliamos a taxonomia com informações semânticas vindas dos glossários e dicionários, inserindo relações de sinonímia.

4.2.1. Pela forma

A intenção foi organizar as entradas em 'famílias', gerando uma melhor organização, ainda que simples, do conhecimento da área. Por isso, buscamos identificar quais entradas seriam 'mães' e quais seriam 'filhos'. Os níveis foram feitos pelo critério lexical. A seguir, apresentamos como se deu o processo de estruturação da lista.

1) Identificamos as entradas que eram unigramas (Entrada Unigrama) e separamos as outras entradas (Entrada Composta). Exemplo: "óleo" seria Entrada Unigrama (EU) e "óleo combustível" seria uma entrada composta (EC).

2) O início de todas as entradas compostas foi analisado, com o intuito de identificar quais delas começavam com uma Entrada Unigrama. Nos casos em que a regra confirmou, a Entrada Composta se tornou filha da Entrada Unigrama.

[*] Deixo aqui meu agradecimento ao Elvis Souza, o programador que ajudou a gerar a taxonomia.

Exemplo: óleo (entrada mãe)

óleo cru (entrada filha)

óleo combustível (entrada filha)

3) Repetimos a regra descrita acima, só que dentro das primeiras famílias já formadas, ou seja, criamos novas famílias dentro da primeira família. Além disso, ao lado de cada entrada 'mãe' passou a aparecer a quantidade de filhos contidos.

Exemplo: óleo (172)

óleo cru (15)

óleo cru aromático (0)

óleo cru asfáltico (0)

óleo cru de base mista (0)

óleo cru de base naftânica (0)

(...)

óleo combustível (14)

4) Incluímos na lista as entradas que não eram unigramas e que não haviam passado na regra ainda. Ou seja, todas as entradas que não receberam atenção. Pelos exemplos, pudemos notar que ainda havia alguns problemas, tais como os casos de plural - "óleos" não entrou na família "óleo" - e os casos de palavras da mesma família, mas sem uma 'mãe', isto é, sem um unigrama na lista – como nos casos das entradas iniciadas com óxido.

Exemplo: óleos pesados (0) (o plural não entrou na família 'óleo')

óleos residuais (0)

óxido de alumínio (0) (óxido não se tornou uma família)

óxido de cálcio (0)

5) Resolvemos a questão do plural nos casos com acréscimo de s/es. Criamos uma condição: se as palavras fossem iguais, com a única diferença de terminar com 's' ou 'es', elas também seriam tratadas pela regra.

Exemplo: óleo (172)

óleo cru (15)
óleo combustível (14)
óleo diesel (14)
óleos (9)
óleos pesados (0)
óleos residuais (0)
(...)

6) Criamos a Entrada Unigrama das famílias que não tinham essa entrada. Uma primeira estratégia foi retirar os artigos “a” e “the” do início de algumas entradas. Outra estratégia foi criar unigramas para entradas que possuíam mais de uma vez “de” ou “a” depois da primeira palavra.

Exemplo: em ‘óxido de alumínio’ temos um sintagma cujo núcleo é seguido por ‘de’. Com relação a esse núcleo, isso pode ser observado mais de uma vez ao longo da lista, como em ‘óxido de cálcio’. Por conta disso, a palavra “óxido” foi trata pela regra e se tornou uma família:

óxido (19)
óxido de alumínio (0)
óxido de cálcio (0)

4.2.2. Pelo sentido

Após estruturarmos os itens da lista pelo núcleo do sintagma nominal, obtivemos uma primeira versão da taxonomia do domínio de óleo e gás, mas ainda com muitos espaços para melhora. Um segundo passo, descrito agora, foi utilizar informação semântica para agrupar os itens. A ideia foi aglutinar, pelo significado, as famílias já formadas, estabelecendo relações entre os possíveis termos. Essa etapa foi realizada através de uma lista de sinônimos, que foi elaborada com base nos glossários utilizados nessa pesquisa. O processo de construção do recurso de sinônimos é explicado a seguir.

4.2.2.1.
A elaboração de um recurso de sinônimos

Os vocabulários utilizados, ainda que aparentemente lineares em seu conteúdo – diferentemente de ontologias/taxonomias –, possuem uma estrutura subjacente, indicada pelas definições (quando existem) ou por indicações presentes em cada entrada. No caso dos glossários, dicionários e tesauros utilizados nesse trabalho, essa estrutura foi considerada e aproveitada, no sentido de trazer maiores informações sobre os termos listados. Ou seja, além de os itens dos vocabulários terem sido selecionados como candidatos a termos da taxonomia, foi feito um trabalho de análise da metalinguagem dos glossários/dicionários/tesauros, com a identificação do que foi sinalizado como sinônimo.

Foi observado que nos cinco glossários havia códigos para expressar relações semânticas, tais como: travessão, parênteses, chaves, colchetes, "ver também", "USE", entre outros. Depois de analisarmos todos os repertórios, bem como todos os códigos possíveis, elencamos, com o fim de elaborar um recurso com relações de sinonímia, os glossários e códigos com relações mais precisas.

Foram quatro vocabulários e quatro tipos de código selecionados, descritos a seguir: (1) no glossário ANP, extraímos as relações marcadas por travessão; (2) no glossário Indústria do Petróleo e Gás, extraímos as relações por travessões e parênteses; (3) no Tesauro 1, extraímos relações de USE e UP (que significa "Use Para") e (4) no Tesauro 2, extraímos as relações de USE e UP.

GLOSSÁRIO	CÓDIGO	EXEMPLO
ANP	Travessão	Programa Exploratório Mínimo – PEM
Indústria do Petróleo e Gás	Travessão	Gás Natural Liquefeito – GNL
	Parênteses	Estocagem Subterrânea de Gás Natural (ESGN)
Tesauro 1	USE	Adubo orgânico USE Fertilizante orgânico
	UP	Acidificação UP Acidulação
Tesauro 2	USE	Cal USE Óxido de cálcio
	UP	Depurador UP Purificador

Quadro 3 – Códigos de sinonímia nos vocabulários

A seguir exemplos de relações de sinonímia extraídas, a partir dos códigos:

Programa Exploratório Mínimo _SINÔNIMO_DE_ PEM
Gás Natural Liquefeito_ SINÔNIMO_DE_ GNL
Adubo orgânico _SINÔNIMO_DE_ Fertilizante orgânico
Depurador _SINÔNIMO_DE_ Purificador

Após extrairmos todas as relações dos vcabulários e códigos selecionados[*], geramos uma lista de candidatos a sinônimos. Essa lista passou por procedimentos de limpeza, assim como as fontes, são eles: (1) eliminação de traduções para o inglês, quando havia; (2) eliminação manual de entradas menos precisas nos códigos travessão e parênteses, e (3) eliminação de informação em parênteses dentro do resultado gerado. Exemplos:

(1) Zirconita #Zircon _SINÔNIMO_DE_ Zircão
(2) Entregue no terminal – DAT
(3) Fonte (Geologia) _SINÔNIMO_DE_ Nascente (Geologia)

Em seguida, os resultados de diferentes glossários e códigos foram unificados como uma lista/recurso de sinônimos e organizados em ordem alfabética.

4.2.2.2. Acréscimo da informação dos glossários

Com o recurso de sinônimos construído, o passo seguinte foi adicionar as relações obtidas na taxonomia. No entanto, antes disso efetivamente ser feito, a lista construída foi inteiramente revisada. A princípio, a lista continha 225 entradas de possíveis sinônimos. Após a revisão, a lista passou a ter 125 relações. A queda se deu por relações repetidas no recurso, vindas dos glossários.

A análise cuidadosa levou a uma distribuição dos pares de sinônimos em três tipos: irmãos gêmeos, irmãos sintagmáticos e irmãos não-sintagmáticos. Como irmãos gêmeos, foram classificados os acrônimos/siglas da lista, por exemplo:

[*] Deixo aqui meu agradecimento à Melissa Costa, programadora que contribuiu com esta tarefa.

OPEP _SINÔNIMO_DE_ Organização dos Países Exportadores de Petróleo

Como irmãos sintagmáticos, foram classificados os sinônimos da lista que são um sintagma nominal e que possuem o mesmo núcleo. Exemplo:

Rocha geradora _SINÔNIMO_DE_ Rocha matriz

Por fim, como irmãos não sintagmáticos, foram classificadas as palavras ou os sintagmas que são sinônimos e que não possuem semelhança lexical. Exemplos:

Navio tanque _SINÔNIMO_DE_ Petroleiro
Declive _SINÔNIMO_DE_ Talude

É válido mencionar que as duplas de irmãos gêmeos – por possuírem total igualdade com o seu par – tornaram-se imediatamente membros da mesma família, como uma entrada só (uma ao lado da outra). Por exemplo, "OPEP" era uma entrada sem família na taxonomia, e por conta do recurso de sinônimos, deixou de ser uma entrada "órfã" e passou a ficar ao lado de "Organização dos Países Exportadores de Petróleo", dentro da família "Organização".

Quanto às duplas de irmãos sintagmáticos – que já compartilhavam, pelo léxico, a mesma família – tornaram-se também uma entrada só, de mesmo modo. Por exemplo, "rocha geradora" e "rocha matriz" já compartilhavam a mesma "mãe" – rocha –, mas eram entradas diferentes na família. Com o acréscimo da sinonímia, elas foram colocadas uma ao lado da outra, como sinônimos.

Isso também valeu para os irmãos não sintagmáticos na maioria dos casos. Por exemplo, toda a família "petroleiro" foi incluída dentro da família mais ampla "navio".

Entretanto, nem todos os casos foram agrupados. A análise manual mostrou que, em alguns casos, o resultado do agrupamento não era uma relação válida, produzindo uma inferência pouco usual. Nesses casos, as famílias permaneceram separadas, embora a relação de sinonímia fosse expressa em ambas as famílias.

Cada relação semântica produzida foi avaliada e classificada tendo em vista a sua **correção** (ou seja, o que foi extraído como sinônimo é, de fato, sinônimo) e a possibilidade de produzir inferências **válidas**. Por exemplo:

casca esférica _SINÔNIMO_DE_ invólucro esférico

Nesse par, temos uma relação de sinonímia válida. Por outro lado, não temos certeza de que, ao agrupar os termos e incluí-los na taxonomia sob o rótulo *casca*, estamos produzindo uma relação válida (uma casca é um invólucro e um invólucro é uma casca, ainda que esta não seja uma maneira usual de perceber esses dois objetos). Com isso, é pouco produtivo, a princípio, que, quem procura por *invólucros à vácuo*, por exemplo, vá encontrá-lo em *casca*, ou que, quem procura por *cascas*, esteja interessado também em *invólucros à vácuo*. Na classificação acerca da validade das inferências, e como estamos em um domínio técnico, optamos por uma posição conservadora, e só consideramos válido o que não causava estranhamento.

	IRMÃOS GÊMEOS	**IRMÃOS SINTAGMÁTICOS**	**IRMÃOS NÃO SINTAGMÁTICOS**
QTD. Total	27	34	64
Correção	27	34	62
Inferência Válida	24	34	43

Quadro 4 – Classificação dos sinônimos da taxonomia

Como podemos observar, na imensa maioria dos casos as relações estão corretas do ponto de vista da relação extraída (98,4%), e apenas 17,6% são responsáveis por inferências que trazem algum estranhamento (como *casca* e *invólucro*). Esse tipo de fenômeno já foi observado em Freitas (2007), e se deve à polissemia.

Como pode ser observado, todos os casos que causaram inferências estranhas vieram dos irmãos não sintagmáticos, isto é, de sinonímia entre palavras distintas. Nossa solução, na taxonomia, para esses casos, consistiu em uma análise "caso a caso". Desta maneira, dois tipos de solução foram tomados: (1) ora

decidimos por manter os sintagmas de uma relação em suas respectivas famílias – não aceitando a inferência –, (2) ora optamos por juntar os pares de sinônimos em uma mesma família. Nos casos em que os pares de sintagmas de uma relação não foram alocados em uma única família, vale destacar, a sinonímia foi indicada duas vezes, nas respectivas famílias.

A análise caso a caso também nos levou uma outra tarefa, para além destas duas soluções: a tarefa de desambiguar famílias com polissemia. Todas as famílias envolvidas com inferência inválida foram analisadas, no sentido de se averiguar se a mãe (o unigrama) possuía mais de um sentido. Essa análise levou em consideração os significados dos filhos. Nos casos em que a polissemia foi constatada, a família se subdividiu: a mãe foi duplicada e passou a alocar somente filhos com o mesmo sentido. Com os exemplos trazidos abaixo, todas as soluções ficam mais claras.

Um exemplo de par de sinônimos que ficou em duas famílias diferentes é "limite de escoamento" e "ponto de ruptura". Analisamos as famílias "limite" e "ponto" e decidimos que juntá-las seria introduzir erro (a palavra "ponto" é altamente polissêmica, o que explica essa dificuldade). A família "limite", por sua vez, continuou como uma família só, com entradas como: *limite de elasticidade*, *limite da área*, *limite de tolerância*, *limite de risco*, *limite de razão gás-óleo*. Já na família "ponto", foi constatada polissemia. Por conta disto, foi dividida em duas: uma família para "ponto", com sentido de *local* e outra para "ponto", com sentido de *medida*.

Na família com sentido de local, colocamos entradas como *ponto de coleta*, *ponto de carga*, *ponto de abastecimento*. Já na família com sentido de medida, colocamos, por exemplo: *ponto de fusão*, *ponto de entupimento*. Quanto à sinonímia, "limite de escoamento; ponto de ruptura" ficou indicada em duas famílias, na família "limite" e na família do "ponto", que se referia a *medida*. Esse é um caso em que estabelecemos uma inclusão dupla, na qual a sinonímia foi indicada em duas famílias.

Agora, um exemplo de par de sinônimos que possibilitou a união de famílias é "cera" e "parafina". Neste caso, toda a família "parafina" pode se encaixar dentro da família "cera" como hipônimos, pois são tipos de *cera*. Além disso, não foi constatada polissemia dentro da família, não havendo, portanto, subdivisões.

Já um exemplo de par de sinonímia que ficou em apenas uma família, mas que gerou desambiguação é "fonte; nascente". A palavra "fonte", assim como

"ponto" também é polissêmica, e por isso separamos os filhos em duas famílias. Na primeira família, temos "fonte" com termos relacionados a água: *fonte termal* e *fonte hidrotermal*. É nessa família que se hospedou o termo "nascente". Na segunda família, temos a família "fonte", como *origem*, exemplos: *fonte de energia*, *fonte de matéria-prima*, *fonte de poluição*.

O caso de "casca esférica; invólucro esférico", mencionado anteriormente, foi solucionado de forma parecida. Ao analisarmos as famílias "casca" e "invólucro", percebemos que o estranhamento derivava de polissemia. A família "casca" ora tinha a ver com *embalagem* e *invólucro*, ora com *comida*. Por isso, separamos "casca" em duas famílias distintas: uma, unida com a família "invólucro", onde ficou: *casca esférica, invólucro esférico* e *casca cilíndrica*. E outra, com os itens relacionados a *comida*: *casca de soja, casca de arroz, casca de amendoim* etc.

No total, 19 relações vindas do recurso de sinônimos – mais especificamente, dos casos de irmãos não sintagmáticos – causaram inferências pouco usuais. Nossa solução para esses casos consistiu em desambiguar todas as famílias das relações, no caso, 38 famílias, e deixar a relação de sinonímia, ora em uma família só, na mais adequada, ora, em duas famílias. Além disso, algumas famílias puderam ser totalmente unidas (como *parafina* e *cera*) e outras permaneceram como distintas. Outras ainda puderam ser, por vezes, ampliadas, como no caso de *fonte*, que possui agora duas entradas na taxonomia. Abaixo, uma tabela com as 19 relações de sinonímia que causaram estranhamento mais as famílias envolvidas, que, por sua vez, passaram pela checagem de polissemia.

FAMÍLIAS		**RELAÇÃO**
1 – Fonte	2 – Nascente	Fonte _SINÔNIMO_DE_ Nascente
3 – Banco	4 – Baixio	Banco de areia _SINÔNIMO_DE_ Baixio
5 – Desvio	6 – Passagem	Desvio; by pass _SINÔNIMO_DE_ Passagem secundária; by pass
7 – Cera	8 – Parafina	Cera de petróleo _SINÔNIMO_DE_ Parafina de petróleo
9 – Criação	10 – Pecuária	Criação de gado _SINÔNIMO_DE_ Pecuária
11 – Casca	12 – Invólucro	Casca esférica _SINÔNIMO_DE_ Invólucro esférico
13 – Cone	14 – Funil	Cone alimentador _SINÔNIMO_DE_ Funil de enchimento

15 – Depósito	16 – Fácies	Depósito miogeossinclinal _SINÔNIMO_DE_ Fácies miogeossinclinal
17 – Ligação	18 – Ponte	Ligação de hidrogênio _SINÔNIMO_DE_ Ponte de hidrogênio
19 – Limite	20 – Ponto	Limite de escoamento; Limite elástico _SINÔNIMO_DE_ Ponto de ruptura
21 – Taxa	22 – Relação	Relação de compressão _SINÔNIMO_DE_ Taxa de compressão
23 – Oleoduto	24 – Ramificação	Oleoduto com derivações _SINÔNIMO_DE_ Ramificação de oleoduto
25 – Alinhamento	26 – Trend	Trend _SINÔNIMO_DE_ Alinhamento
27 – Solução	28 – Água	Água mãe_ SINÔNIMO_DE_ Solução mãe
29 – Dispositivo	30 – Mecanismo	Dispositivo _SINÔNIMO_DE_ Mecanismo
31 – Camada	32 – Nível	Camada dissimuladora _SINÔNIMO_DE_ Nível mascarador
33 – Processamento	34 – Processo	Processamento em lote _SINÔNIMO_DE_ Processo em batch
35 – Teste	36 – Ensaio	Teste de metais _SINÔNIMO_DE_ Ensaios de metais
37 – Tubo	38 – Cabo	Tubo de óleo de freio _SINÔNIMO_DE_ Cabo de óleo de freio

Quadro 5 – Relações de famílias polissêmicas

5
Resultados

5.1.
Apresentando e interpretando os resultados

Após a primeira tentativa de estruturação da lista, pela forma, descrita acima, percebemos, em primeiro lugar, que o tamanho aparente da lista havia encurtado de forma significativa. O número de entradas, originalmente, antes de uma estruturação hierárquica, era 29.035. Depois que agrupamos os itens em famílias, se não considerarmos os filhos (ou seja, se olharmos apenas para as entradas desconectadas e as mães), a lista passou a apresentar 7.896 entradas, uma diminuição do tamanho aparente em mais de 60%. Apesar de o tamanho, a princípio, parecer menor, considerando também os filhos, em função dos procedimentos realizados, a lista passou a conter um total de **30.626** entradas.

Das 30.626 entradas totais da lista estruturada, 22.730 entradas são filhas, ou seja, são entradas/sintagmas compostos relacionados ou encaixados em alguma palavra. Isso significa que a maior parte da lista de candidatas a entidade são sintagmas compostos: em torno de 74,2% da lista. Além disso, constatamos, no total, 4.311 mães na taxonomia. Essas mães podem ser de dois tipos: tanto podem ser os unigramas na lista, funcionando como os nodos principais, quanto podem ser as filhas de unigramas, que também são mães. Exemplo:

óleo (172)

    óleo cru (15)

    óleo cru aromático (0)

No caso acima, temos que “óleo” é uma entrada mãe-unigrama, “óleo cru” é tanto entrada filha de “óleo” como entrada mãe de “óleo cru aromático” e, por fim, “óleo cru aromático” é uma entrada filha de “óleo cru” e de “óleo”. As mães-unigramas da estrutura consistem em 2.626 entradas, já as mães que são também

filhas constituem 1.685 entradas. De modo geral, as entradas mães-unigramas (nodos) representam apenas 8,5% da lista completa.

A taxonomia possui ainda 5.270 entradas desconectadas, isto é, que não pertencem a nenhuma família. Tal número indica 17,2% da lista inteira de 30.626 entradas. Essas entradas não foram consideradas unigramas ou sintagmas compostos da taxonomia, mas sim termos sem família. Isso significa que os unigramas são representados apenas pelas mães-unigramas – os nodos da estrutura, que permitem encaixes –, enquanto os termos compostos significam aqueles que formalmente se inseriram em algum unigrama, os membros da família. A ideia é que, com procedimentos futuros, os termos desconectados possam ser revistos, de modo a se encaixarem semanticamente ou gerarem novas famílias, ou ainda serem eliminados. Abaixo, seguem os dados mencionados em tabelas e, em seguida, uma lista das cinquenta entradas mais produtivas, isto é, com mais filhos.

TIPO DE ENTRADA	QUANTIDADE	VALOR RELATIVO
Termos MÃES-unigramas	2.626	8,5%
Termos FILHOS (sintagmas compostos)	22.730	74,2%
Termos SEM FAMÍLIA (entradas desconectadas)	5.270	17,2 %

Quadro 6 – Análise quantitativa dos resultados

processo (276)	bomba (73) formação (73)
sistema (195)	fator (72) fluxo (72) tipo (72) velocidade (72)
análise (172) óleo (172)	equação (70) número (70) perfuração (70)
modelo (152)	material (69)
método (131)	produção (68)
valor (128)	ensaio (67)
pressão (118)	estrutura (66) fase (66) razão (66)
depósito (117)	agente (62)
perfil (112)	camada (61) força (61) propriedade (61)
poço (108)	linha (60) superfície (60) taxa (60)
gás (102)	controle (59) escoamento (59) log (59)
mapa (99)	levantamento (56) região (56) tensão (56)
efeito (96)	condição (54) dado (54) energia (54)
válvula (93)	relação (53)
filtro (92)	perfilagem (52) solução (52) tubo (52)

tempo (89)	amostra (51) corrente (51) teor (51)
equipamento (88) ponto (88)	atividade (50) custo (50) motor (50)
unidade (87)	quantidade (50) reação (50) volume (50)
concentração (85) onda (85) temperatura (85)	ácido (50) mistura (49) índice (49)
água (85) curva (81) projeto (81)	coeficiente (48)
fluido (80)	campo (47) capacidade (47) falha (47)
área (77)	fonte (46) lama (46) resistência (46)
diagrafia (76)	medidor (45) padrão (45) resultado (45)
teste (75)	bacia (44) parâmetro (44)
função (74) rocha (74) zona (74)	aumento (43) carga (43) tanque (43)

Quadro 7 – As cinquenta “mães” mais produtivas

A primeira coisa que pontuamos é que nas cinquenta entradas com mais filhos, que considera também as entradas com a mesma quantidade de filhos, a palavra petróleo não apareceu, nem combustível ou gasolina, palavras que possivelmente qualquer pessoa com conhecimento raso no assunto relacionaria ao domínio. Ao mesmo tempo, algumas palavras já esperadas – óleo, gás, poço – apareceram, ainda que nem tanto no topo da estrutura.

Percebe-se também a presença de palavras nem tanto do domínio de óleo e gás, mas sim, de uma maneira geral, da área técnica/acadêmica: análise, modelo, método, teste, fator, equação, resultado, aumento, coeficiente, etc. Isso é compreensível, já que o domínio é bastante técnico e o *corpus* também pertence a esse universo, uma vez que é constituído por texto acadêmico. Ainda assim é interessante perceber que o método é sensível aos gêneros utilizados. Em outros gêneros, uma taxonomia de óleo e gás pode apresentar um resultado bastante diferente.

É válido mencionar que a palavra ‘tempo’ apareceu entre as vinte com mais filhos, o que talvez enfatize que tempo é algo muito relevante dentro da área. O tempo de vida de um poço, o tempo de produção, transportação, o tempo de certos processos. Por ser uma área que envolve muitos procedimentos e processos, e até mesmo dinheiro, informação sobre tempo pode ter um valor enorme. Isso parece ser novamente ratificado porque a palavra ‘processo’ é a primeira da lista. Todas as nossas fontes geraram 276 diferentes tipos de processo que estão conectados ao domínio de óleo e gás, o que mais uma vez sugere a importância do tempo e da ação.

Um dado complementar são as palavras que, apesar de serem substantivos, estão ligados a uma atividade, algo próprio da classe gramatical verbo: análise, formação, concentração, aumento, controle, produção, processo, perfuração, escoamento, levantamento, perfilagem. Há muitas nominalizações na lista. Todos esses indícios nos levam a crer que as entidades (ou os termos) no domínio de óleo e gás podem ter caráter mais dinâmico. O fator geográfico também teve destaque em entradas como mapa, unidade, área, região, campo, zona, bacia, entre outros. A princípio, esse resultado nos leva a pensar que locais/ambientes, tempo e processos parecem ser categorias relevantes dentro da área.

Curiosamente, as dez palavras (não entradas) mais frequentes da lista dos glossários e da lista gerada pelo TermoStat estão entre as cinquenta mães com mais filhos. Para os glossários, as palavras são: gás, poço, óleo, pressão, perfuração, produção, água, processo, onda e válvula. Para o TermoStat, as palavras são: processo, modelo, óleo, valor, sistema, gás, água, concentração, análise e método. Quanto às dez palavras mais frequentes da lista de palavras-chave, seis delas são pais que estão entre cinquenta termos com mais filhos: análise, gás, óleo, método, controle e energia. Faz sentido, já que a lista de palavras-chave é a menor lista utilizada, tendo provavelmente, menos palavras.

5.2. Provendo melhorias na taxonomia

Com a lista de palavras estruturada, o que facilita bastante a visualização, foi mais fácil perceber erros sistemáticos (e maneiras de eliminá-los). Por exemplo, percebemos que algumas famílias possuíam como "mãe" palavras inapropriadas, vindas de classes como pronome e preposição. Com relação aos procedimentos que envolvem sentido, percebemos, durante o processo, que o problema da polissemia não era algo restrito às famílias que vieram do recurso de sinônimos, mas algo que percorria a taxonomia inteira, sobretudo as famílias com mais integrantes.

Além disso, também pudemos detectar entradas compostas, cujos modificadores (adjetivos), apesar de muito presentes em textos acadêmicos, não serviam como candidatos a termos da taxonomia. Muitos, inclusive, pareceram derivar de erros de análise gramatical (POS), vindos do TermoStat.

Nesta seção, pretendemos apresentar as soluções tomadas para os principais erros encontrados no resultado. O objetivo dessas medidas foi o de aprimorar a taxonomia.

5.2.1. Uma leitura distante de erros

Assim como outros trabalhos de identificação automática de termos, que mencionam a presença de elementos indesejados nos termos extraídos (cf. capítulo 3), detectamos diversos tipos de ruído na taxonomia e descrevemos aqui como solucionamos os erros mais gerais. De maneira geral, foi realizada uma leitura distante da taxonomia como um todo. Isso significa que tanto as mães-unigramas, como os filhos e as entradas desconectadas foram brevemente analisados, com o intuito de se detectar imprecisões.

Quanto aos unigramas, buscou-se identificar os que eram iniciados por preposições, pronomes e artigos, ou por qualquer elemento que nitidamente representasse erro. Os casos encontrados foram descartados, mas é justo afirmar que nem todos os filhos presentes dentro dessas famílias foram, sem leitura, eliminados, poucos deles foram realocados em outra família. Por exemplo, na família "by", o filho "by pass" passou a pertencer a família "desvio".

A taxonomia, em seu topo, apresenta as maiores famílias, isto é, os termos com mais filhos. Os dados vão se apresentando por quantidade de filhos e ordem alfabética até que se chegue a uma imensa cauda longa de termos sem família, ou seja, de entradas órfãs. Com o objetivo de melhorar a qualidade dos termos apresentados na taxonomia, essa cauda longa de termos desconectados foi analisada brevemente. A ideia aqui era, a princípio, investigar se a cauda longa representava itens relevantes para o domínio em questão ou se eram, na maioria, ruído.

Impressionantemente, muitas entradas são dignas do estatuto de termo na taxonomia. Muitas delas estão desconectadas pelo simples fato de que não houve coincidência formal com os unigramas. Por exemplo, entradas como "amido", "éster", "sienito" e "zirconita" não possuem mãe ou filhos, pelo critério formal/lexical, com os outros itens da taxonomia. No entanto, poderiam entrar, por um critério semântico, em famílias como "polímero", "composto", "rocha" e "mineral", respectivamente. Além disso, há muitas siglas, que vieram do siglário (e

não dos glossários), que recaíram sobre a cauda longa, mas que podem, e precisam, ser realocadas em famílias.

Dada a situação, a partir da leitura distante, o que se pretendeu foi manter o que parecia relevante e eliminar o que parecia, de fato, erro. Uma estratégia para a eliminação foi pesquisar por elementos mórficos que indicassem adjetivos e verbos, por exemplo, palavras terminadas em ado/ada. A partir daí também surgiu a limpeza de entradas iniciadas por determinados prefixos. Conseguimos detectar alguns prefixos que impediram que certos termos fossem alocados dentro da família certa. Por exemplo, o termo "supernavio-tanque" estava órfão na taxonomia, mas, por meio desta análise, conseguimos realocá-lo para a família "navio".

Detectamos três prefixos que geraram ruído na taxonomia: *super*, *bio* e *micro*. Os casos encontrados na cauda longa da taxonomia foram avaliados manualmente e, quando, puderam ser alocados em alguma família – como em "supernavio-tanque" –, foram. Com relação ao prefixo *bio*, "bioetanol" e "biometano" passaram a morar em "etanol" e "metano", respectivamente, por exemplo. Já com relação a *micro*, temos que "microrreator" e "microfácies" foram alocados em "reator" e "fácies".

Pequenas famílias também foram realocadas durante este processo, por exemplo, a família de "biodiesel" (biodiesel de palma, biodiesel de soja, biodiesel de gordura, etc) passou a ficar dentro da família "diesel", tal como a família "microorganismo" (microorganismo aeróbico, microorganismo anaeróbico, microorganismo do solo, etc) foi para "organismo".

A maioria dos casos de termos órfãos, que apresentavam os prefixos acima, puderam ser colocados dentro de uma família. Dos 20 casos de prefixo *super*, 13 puderam ser realocados, 1 continuou órfão e 6 eram casos de erros – adjetivos, por exemplo: supervisório, superior, entre outros. Dos 36 casos de *bio*, 22 foram realocados, 8 permaneceram sem família e 6 eram erros. Sem contar as 7 pequenas famílias que também foram realocadas. Por fim, dos 36 casos de *micro*, 27 termos foram realocados, 6 ficaram órfãos na cauda longa e 3 eram erros. Além disso, uma família inteira também pode ser reencaixada em outra.

Convém lembrar que a eliminação de ruído por leitura distante ou pelas pesquisas semiautomáticas não foi suficiente para solucionar, no âmbito deste trabalho, todos os problemas das entradas desconectadas e dos unigramas. É imprescindível que se encontre estratégias para realocar os termos relevantes que

ficaram soltos e as mães genéricas que não foram identificadas. De mais a mais, para além da cauda longa e dos unigramas, também as entradas filhas foram analisadas. A próxima subseção trata do que foi feito com elas.

5.2.2. Filtro de modificadores

Ao longo do processo de construção da taxonomia, identificamos termos compostos, cujo modificador não era, de fato, relevante para o domínio do óleo e gás, mas sim comum no gênero acadêmico ou fruto de erro de análise gramatical. Elencamos algumas palavras que, de maneira geral, funcionaram como filtros para eliminarmos entradas inapropriadas ao longo da taxonomia. Eles foram selecionados manualmente, durante o processo de verificação de polissemia em famílias, e, em seguida, utilizados para encontrar e excluir entradas ruins. Nesta seção, vamos brevemente apresentá-los.

Um caso bastante comum foi o de entradas cujo modificador era um verbo na forma participial. Por exemplo, as entradas compostas "teste realizado", "água utilizada", "solução obtida", "mecanismo proposto" e "ponto adotado" não são bons candidatos a termos de uma terminologia do óleo e gás, mas faziam parte da taxonomia. Muito provavelmente, apareceram na lista não como consequência de fontes mais confiáveis, como glossários, mas da análise feita pelo TermoStat, que deve ter considerado todas as formas como *adjetivo*.

Outro tipo comum de erro foram as entradas cujo modificador estava incompleto ou representava algum tipo de erro de análise. Por exemplo, as entradas "cabo de são" e "cabo de santa" estão erradas porque estão incompletas. Assim, entradas nas quais os modificadores eram desse tipo foram eliminadas.

Já nas entradas "processo tabela" ou "teste figura", percebemos que houve erro de análise, provavelmente, proveniente da ferramenta TermoStat. Vale lembrar que não pudemos trabalhar com um *corpus* bem processado, tampouco com ele morfossintaticamente anotado, o que teria auxiliado na detecção mais precisa de termos compostos corretos. O material que usamos – e que foi incluído na ferramenta TermoStat –, é uma versão preliminar do *corpus*, derivada da conversão de documentos PDF para o TXT, sem qualquer tratamento especial para figuras, tabelas e gráficos, que são elementos bastante frequentes em teses e dissertações.

Também são frequentes itemizações, que levam a "modificadores" que são letras soltas: "camada b", "ponto c" e "teste f" não parecem relevantes de domínio, mas consequência do gênero, e, por isso, foram eliminadas. Por fim, assim como em outros trabalhos, modificadores muito gerais também foram elencados como filtros. Alguns exemplos são "ótimo", "exata" e "ultra", que aparecem em entradas como: "ponto ótimo", "solução exata", "água ultra". O anexo 1 lista os modificadores utilizados.

5.2.3. Casos de coordenação

O problema de entradas com coordenação, mencionado no GENIA (KIM, 2003) e em Wendt et al. (2010) também apareceu na taxonomia, especialmente nos filhos (termos compostos) e na cauda longa. Todos os casos foram tratados individualmente. De maneira geral, quatro formas de solução foram tomadas:

(1) A coordenação se manteve no termo.

(2) Os termos com coordenação viraram termos diferentes na taxonomia.

(3) A parte coordenada, ou mais partes, foi eliminada da entrada.

(4) A parte coordenada virou um sinônimo da primeira parte.

A primeira foi a solução dada para entradas que significavam acrônimos ou organizações, empresas. Exemplos: "agência nacional do petróleo, gás natural e biocombustíveis" (ANP) e "ministério de minas e energia". A segunda foi a solução dada para as entradas que podiam representar dois ou mais termos, por exemplo: "campo de petróleo ou de gás natural", "óleos e gorduras residuais" e "terminal marítimo, fluvial ou lacustre".

Já a terceira solução foi para casos de erro ou sobra nas entradas, tais como: "óleo lubrificante acabado envasado e a granel" – que ficou apenas "óleo lubrificante" – e "pressão 1500 psig e acima" – que ficou "pressão 1500 psig". Por fim, a quarta e última solução foi usada nos casos de coordenação com "ou", em entradas como "lavra ou produção", que virou "lavra; produção" e "óleo cru ou bruto", que virou "óleo cru; óleo bruto".

Em torno de 250 casos de coordenação foram encontrados avaliados. Cabe destacar aqui que as soluções (1) e (2) foram as mais utilizadas. Foi observado, no que concerne à primeira solução, que alguns casos que, a priori, pareciam erros, na

verdade eram acertos. Os quatro casos abaixo, a princípio, pareciam casos de entradas com coordenação que precisavam de separação. No entanto, a análise mostrou que, apesar de não virem na forma clássica de bigrama ou trigrama sem coordenação, essas entradas representavam uma unidade dentro do domínio de óleo e gás, possuindo, inclusive, acrônimos.

1) custo, seguro e frete = CIF
2) desancoragem, movimentação e ancoragem = DMA
3) qualidade, saúde, meio ambiente e segurança = QSMS
4) teor de óleo e graxa = TOG

Além disso, nem todos os casos da solução (2) representaram apenas a separação dos itens coordenados. Muitos sintagmas, além de serem subdivididos, também permaneceram na forma maior. Por exemplo, em "plano de inspeção e manutenção dos equipamentos", o sintagma coordenado "manutenção de equipamentos" foi adicionado à família "manutenção", mas o sintagma completo "plano de inspeção e manutenção dos equipamentos" não foi desfeito, ao contrário, foi mantido em "plano". O mesmo ocorreu em "análise de perigos e falhas operacionais", tanto o sintagma com coordenação quanto os itens da coordenação foram conferidos na taxonomia.

Por fim, também foram encontrados casos de erros nas entradas com coordenação. Por exemplo, os sintagmas "discordância ou não concordância" e "resultado e discussão" são típicos do gênero acadêmicos, mas pouco informativos sobre o domínio do óleo e gás. Esses casos foram eliminados.

5.2.4. OpenWordNet.PT

Não foi possível detectar e resolver manualmente, no âmbito desta dissertação, todos os casos de polissemia de todas as famílias presentes na taxonomia. Entretanto, demos início a este processo com as 38 famílias. Cabe também destacar que utilizamos a OpenWordNet-PT[7] – uma versão alinhada com a WordNet de Princeton para o português – como recurso de auxílio na tarefa de desambiguação de famílias, e essa é uma ideia que pode ser explorada no futuro.

[7] http://wn.mybluemix.net/

Além de consultarmos nela as possibilidades de sentido dos termos mães-de-família, também utilizamos, por vezes, os códigos para a criação de novas famílias, quando era o caso.

Um exemplo é o caso da família “processamento” (quadro 8), que fez parte da leva de entradas que foram avaliadas quanto à polissemia. Depois da checagem dos filhos – e com base na OpenWordNet-PT – decidimos subdividir a família em duas: uma guardando o sentido que vem da ciência da computação, e, outra, ligada a ideia de procedimento.

SENTIDO	EXPLICAÇÃO	EXEMPLO
PROCESSAMENTO 1 13455487-n data_processing | processamento de dados	((computer science) a series of operations on data by a computer in order to retrieve or transform or classify information)	processamento de dados processamento de imagens processamento sísmico
PROCESSAMENTO 2 13541167-n processing | processamento	(preparing or putting through a prescribed procedure; "the processing of newly arrived immigrants"; "the processing of ore to obtain minerals")	processamento de óleo processamento de petróleo processamento de gás natural

Quadro 8 – Desambiguando a entrada “processamento”

SENTIDO	EXPLICAÇÃO	EXEMPLO
CASCA | INVÓLUCRO 1 04605726-n wrapping, wrapper, wrap | embalagens, invólucro, cobertura, capa	(the covering (usually paper or cellophane) in which something is wrap)	casca esférica invólucro esférico casca cilíndrica
CASCA 2 11683556-n shell | casca	(the hard usually fibrous outer layer of some fruits especially nuts)	casca de soja casca de noz casca de amendoim

Quadro 9 – Desambiguando a entrada “casca”

No caso do par “casca” e invólucro” (quadro 9), que não envolve apenas desambiguar uma família, mas a junção de duas, o recurso serviu, por exemplo,

como uma orientação sobre como unir e separar os conceitos. Como foi mencionado anteriormente (seção 4.2.2.2), depois da análise, foram deixadas duas entradas na taxonomia, uma com sentido de embalagem, e outra, ligada ao sentido da casca de alimentos.

5.3. Novos resultados

Depois dos procedimentos de apuramento da estrutura, observamos que pouca coisa mudou quanto à ordem das cinquenta mães com mais filhos na taxonomia. De fato, a maioria das famílias que ocupavam esta posição permaneceram como produtivas, mesmo após limpezas. No entanto, quatorze delas deixaram o posto: *tipo, relação, solução, amostra, custo, quantidade, mistura, campo, fonte, resistência, resultado, parâmetro, carga e tanque.*

Isso pode ter acontecido devido a cortes de filhos derivados de modificadores genéricos, por exemplo. Outra possível razão é a eliminação de mães impróprias, tal como *tipo*. Além disso, as famílias "solução", "fonte" e "relação", por exemplos, foram avaliadas e lapidadas quanto à polissemia. Com a eliminação ou com a diminuição do número de filhos dessas quatorze entradas, outras duas famílias passaram a compor o rank de mais produtivas: *contrato* e *camada*. Abaixo, é possível visualizar a nova apuração das cinquenta mães com mais filhos.

processo (259)	velocidade (68) número (68)
sistema (191)	equação (67)
análise (163)	produção (66) estrutura (66)
óleo (158)	água (65)
modelo (134)	razão (64)
método (120)	temperatura (63) fator (63)
depósito (115)	área (62) dado (62)
ensaio; teste (112)	material (61) ponto (61)
perfil (107)	escoamento (60) agente (60)
poço (106)	controle (59)
pressão (100)	força (58) propriedade (58) log (58)
mapa (98)	taxa (57)
efeito (94)	levantamento (56)
gás (93)	superfície (55)

válvula (91) filtro (91)	linha (54)
valor (88)	fase (53)
unidade (87)	teor (52) perfilagem (52)
equipamento (84) onda (84)	contrato (51) tubo (51)
projeto (81)	camada (50) motor (50) ácido (50) tensão (50)
concentração (76) diagrafia (76)	corrente (49) energia (49) reação (49)
tempo (75) zona (75) fluido (75)	índice (48) coeficiente (48)
rocha (74)	atividade (47) falha (47) região (47)
curva (73) função (73) formação (73)	volume (46) capacidade (46) condição (46)
bomba (71) fluxo (71)	lama (46) medidor (45) padrão (45)
perfuração (70)	bacia (44) aumento (44)

Quadro 10 – As novas cinquenta “mães” mais produtivas

Partindo do fato de que a maioria das famílias permaneceram as mesmas, é possível ainda se valer das primeiras interpretações. Palavras como “processo”, “tempo”, “velocidade”, “produção”, “contrato”, “atividade” e “capacidade” validam a leitura de que o domínio do óleo e gás tem o tempo e a produtividade como fatores primordiais. Ainda como adicional, as nominalizações – “análise”, “formação”, “produção” e “aumento”, por exemplos – mais uma vez podem ser um sinal de que os termos (e entidades) da área de óleo e gás podem ter um caráter mais dinâmico, e menos estático, como seria próprio do substantivo.

Aliás, sobre isso, também nos termos compostos, nos modificadores, essa natureza mais verbal se apresenta. É possível encontrar diversas formas participiais utilizadas como adjetivos. Por exemplo, na família *gás*, temos termos compostos como: *gás associado*, *gás condensado*, *gás liquefeito*, *gás comprimido*, *gás combinado* etc. Já na família *poço*, temos: *poço exploratório*, *poço abandonado*, *poço ativo*, *poço tamponado*. Essa é uma característica própria dos sintagmas compostos da taxonomia e pode ser observada em diversas outras famílias, tais como *óleo* e *água*, por exemplos. Tudo isso corrobora para a ideia da relevância do tempo e da ação.

Sobre uma avaliação mais quantitativa, percebemos que a taxonomia diminuiu de 30.626 entradas totais para **27.794** itens. Dentre estes, **20.786** são entradas filhas, isto é, são os sintagmas compostos da taxonomia, que ocupam 74,7% da estrutura. Quanto às entradas mães, se considerarmos os dois tipos, explicados na seção 5.1, temos 3.949 valores. No entanto, como unigramas da taxonomia, temos apenas **2.438** mães, que por sua vez, representam 8% de todas as

entradas da estrutura. Por fim, com relação aos itens desconectados, que ficaram sem família, eles correspondem a **4.570** itens da estrutura, o que, por sua vez, corresponde a 16, 4% da taxonomia.

É justo observar que os unigramas gerados pelas ferramentas AntConc e TermoStat foram utilizados em uma análise comparativa com os unigramas presentes na taxonomia. Isto é, comparamos as entradas mães-unigramas da nossa estrutura com os unigramas que o AntConc e o TermoStat geraram em comum. Sem delongas, entre ambas as listas de unigramas, houve 1.263 entradas compartilhadas, o que significa 29,8% das listas unidas. O número não chega a ser expressivo, mas vale lembrar que a lista de unigramas comuns entre as duas ferramentas possui 6.020 itens, enquanto a taxonomia possui apenas 2.438 unigramas. Logo, é natural que a maioria dos dados do AntConc e TermoStat não correspondam a dados da taxonomia.

Além disso, os resultados das ferramentas consideraram diversas classes gramaticais e não passaram por nenhum tipo de limpeza, como ocorreu com a taxonomia. Dessa forma, os resultados dessa exploração atendem às expectativas possíveis e 1.263 unigramas compartilhados, considerando o número de unigramas da taxonomia, significa um pouco mais da metade de termos simples comuns.

Pós-limpeza	
TIPO DE ENTRADA	QUANTIDADE
Termos MÃES-unigramas (unigramas)	2.438 (8%)
Termos FILHOS (sintagmas compostos)	20.786 (74,7%)
Termos SEM FAMÍLIA (entradas desconectadas)	4.470 (16,4%)
Pré-limpeza	
TIPO DE ENTRADA	QUANTIDADE
Termos MÃES-unigramas (unigramas)	2.626 (8,5 %)
Termos FILHOS (sintagmas compostos)	22.730 (74,2 %)
Termos SEM FAMÍLIA (entradas desconectadas)	5.270 (17,2%)

Quadro 11 – Comparação dos resultados

Como podemos observar pelas tabelas, apesar de que centenas de famílias inteiras foram eliminadas semi-automaticamente, bem como entradas filhas e desconectadas impróprias, em termos de valor relativo, as limpezas não provocaram um impacto impressionante nos números. Isto é, a porcentagem apresentou uma leve diminuição, em comparação com o resultado anterior às limpezas. No entanto, cabe destacar que todas as quantidades foram, ainda que pouco, diminuídas.

Além disso, também houve redução no tamanho aparente da taxonomia. Como a taxonomia foi construída de forma similar a uma árvore de diretórios, inicialmente, o que pode ser observado na estrutura são as mães-unigramas, os nodos, e os termos desconectados. Os filhos só são vistos na medida em que se navega dentro de uma família da estrutura. Em um primeiro momento, a terminologia de 29.035 entradas passou a ser vista com apenas 7.896 entradas, depois da estruturação hierárquica. Após as limpezas, essa visualização passou para 7.008 itens. É possível visualizar a interface da taxonomia na imagem abaixo:

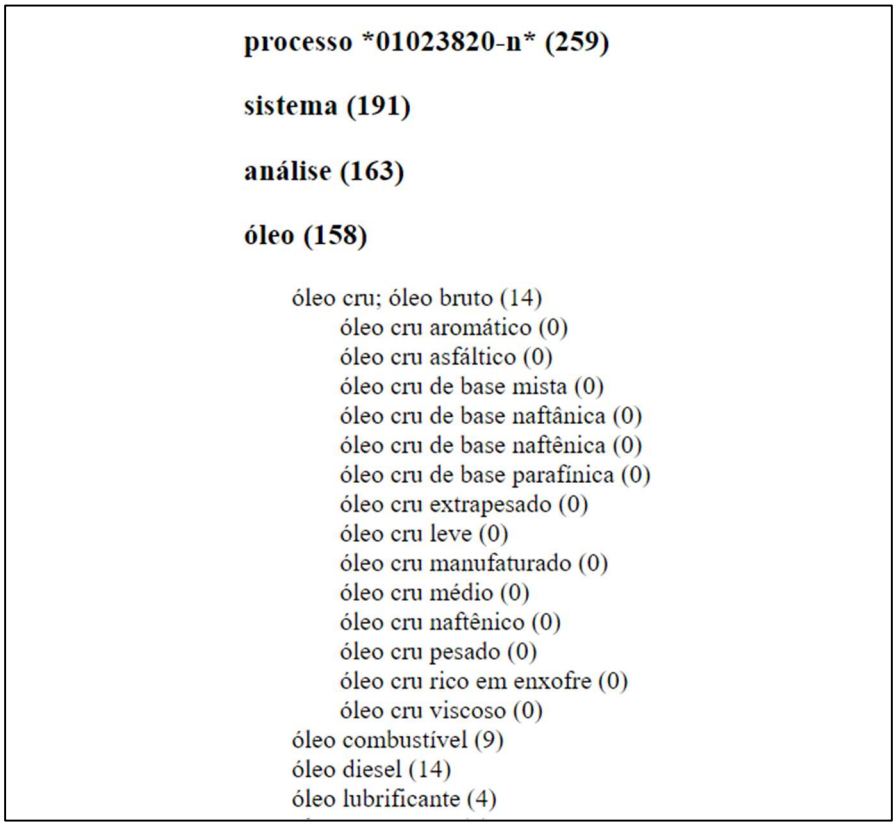

Figura 4 – Interface da 1ª versão da taxonomia de óleo e gás

6
Considerações finais

Nas páginas iniciais da dissertação, declaramos que o presente trabalho possuía diversos objetivos, por exemplo: (1) ajudar a suprir a falta de recursos linguísticos de domínio no PLN e (2) oferecer um recurso que auxilie a organizar o conhecimento de uma área e na extração de informação de um setor industrial. Ao longo deste trabalho, um processo de construção semi-automática de uma taxonomia voltada para o domínio de óleo e gás foi descrito. Como produto da pesquisa, apresentamos a primeira versão de uma taxonomia do óleo e gás. Com isso, é possível dizer que os objetivos um e dois foram contemplados.

Além disso, outra motivação deste trabalho é que a taxonomia pudesse ser utilizada como passo inicial para a anotação de entidades mencionadas de domínio especializado. Uma vez que partimos da obtenção automática de terminologia para a construção de uma estrutura de conhecimento, era importante frisar que a obtenção de termos de domínio representava também um suporte para o reconhecimento de entidades.

Sobre isso, cabe novamente mencionar que, no contexto do projeto BIG OIL, a taxonomia apresentada já vem sendo usada como subsídio para a elaboração de um *tagset* de entidades mencionadas do domínio. Isso significa que a taxonomia vem sendo utilizada como ponto de partida para a identificação e classificação das entidades mencionadas de um *corpus* de óleo e gás, composto por em torno de quatro mil documentos acadêmicos. Assim, um outro objetivo se concretiza.

Ademais, demonstrando que dos termos, podemos chegar às entidades, pudemos promover uma articulação entre uma área aplicada da linguística e uma aplicação do PLN: terminologia e entidades mencionadas. Essa relação entre as áreas, como pode ser percebido no trabalho, não é uma novidade. No entanto, aqui, ela teve uma oportunidade de ser formalmente apresentada e discutida. Acreditamos na importância da construção de diálogos como este, especialmente, tendo em vista o distanciamento que há entre PLN e linguística. Tal distanciamento é tolo, uma vez que tanto o PLN pode se beneficiar no conhecimento linguístico, como a linguística pode aprender a lidar e a olhar para a língua, através do PLN.

Com relação à metodologia da pesquisa, este trabalho segue com uma identidade própria a forma descrita por Bordea et al (2015). Segundo os autores, a construção de uma taxonomia implica (1) extração dos termos (2) descoberta de relações entre os termos e (3) construção da taxonomia. No presente trabalho, os passos (2) e (3) foram realizados simultaneamente. Ou seja, as relações foram detectadas, na medida em que construímos uma hierarquia entre os termos, ou melhor, justamente porque estruturamos os termos. Diferentemente dos sistemas do SemEval, não estávamos focados na utilização dos melhores métodos para se obter as melhores relações e ou hiperônimos. Não partimos do *corpus*, de padrões linguísticos ou métodos estatísticos – em grande parte, porque não foi possível: não tínhamos um *corpus* bem processado e anotado.

O que tínhamos era a terminologia obtida pela ferramenta TermoStat, os glossários e as palavras-chave dos documentos acadêmicos. Então, na realidade, nossas relações taxonômicas emergiram, em um primeiro momento, devido a um critério formal. Apesar de que só o critério formal parece pouco para orientar uma taxonomia – e desejamos impor esforços para ampliar essa realidade –, ainda assim, é possível dizer que foi uma maneira interessante de se obter as hiperonímias e, além de tudo, por causa desta ideia, o que era uma lista linear, pode ganhar uma nova visualização hierárquica, o que, por sua vez, possibilitou um olhar totalmente novo para os dados.

Quanto aos dados, a interpretação de que o domínio possui, como um todo, termos e modificadores mais dinâmicos é interessante, já que, como colocam Krieger e Finatto (2004), os termos possuem, geralmente, um caráter mais nominal, ou seja, estático. Isso aponta para a necessidade de produtividade dentro do setor. Ao mesmo tempo, é possível validar com mais precisão esta leitura, olhando para padrões linguísticos nos tipos de nominalizações realizadas e calculando os casos. Agora, ainda sobre a taxonomia proposta, deve-se dizer que é imprescindível avançarmos na melhoria de seu desenvolvimento.

Um grande ponto a melhorar seria a cauda longa. Nela, temos diversos tipos de ruído convivendo com diversas entradas que merecem o estatuto de termo. Portanto, seria oportuno que fosse realizada uma lapidação mais profunda, não a olho nu de uma única pesquisadora. Afinal, frequência é algo relativo e o que sobrou não necessariamente é lixo. O que for, merece sair, mas o que pode ser estruturado, merece ser aproveitado. Afinal, o critério que embasou a estruturação foi lexical,

pelo núcleo, porém, se tivéssemos estruturado a lista por grupos semânticos ou por outra parte das palavras – como sufixos –, a cauda longa poderia ser bem diferente. Por exemplo, no domínio de óleo e gás, temos que palavras terminadas com o sufixo "ita" são tipos de rocha. Futuramente, pretendemos agrupar essas palavras que ficaram desconectadas, em função do agrupamento pelo núcleo. É justo mencionar que linguistas do projeto BIG OIL já começaram a olhar para a cauda longa, mas o trabalho realizado não pôde ser aproveitado a tempo desta dissertação.

Outros pontos importantes a melhorar – que, inclusive, afetam a cauda longa – incluem a (1) utilização de mais conhecimento semântico para a estruturação e agrupamento da lista, o que geraria uma estruturação ainda mais enxuta e hierárquica do domínio, e (2) a utilização de mais anotação linguística, tal como informação sintática, para resolver problemas oriundos de análise gramatical incorreta e para lidar melhor com questões de plural e singular.

Com relação ao primeiro ponto, seria interessante, por exemplo, a utilização de outras informações oriundas dos glossários por meio dos códigos. Apenas alguns códigos presentes nos glossários foram utilizados para ampliar a informação presente na taxonomia, tais como USE e UP, que serviram para a inclusão de sinonímia. Além destes, poderiam ser incluídos novos, por exemplo, o TR, que significa termos relacionados. De mais a mais, outro tópico seria a desambiguação completa da taxonomia, que, manualmente, não pôde ser totalmente realizada. Ou seja, também seria de grande melhora se todos os casos de polissemia fossem analisados e solucionados, igual às 38 famílias verificadas.

Adicionalmente, também é possível pensar na utilização do *corpus* Petrolês, e, até mesmo, em sua versão maior, para a identificação de relações taxonômicas na área de O&G. Esse tipo de pesquisa foi abordado ao longo da dissertação, por meio de Hearst (1992), que identificou padrões linguísticos para se extrair relações de hiperonímia de forma automática em *corpus*. Outra forma de pensar na melhoria da taxonomia, quanto à utilização de informação semântica, seria adaptando o trabalho de Hearst (1992) para o universo desta pesquisa. Com isso, por meio de *corpus* – e uma abordagem linguisticamente motivada – poderíamos identificar novas relações hierárquicas entre os termos da taxonomia. Em contrapartida, para isso, seria necessária a anotação sintática.

Neste trabalho, utilizamos a ferramenta TermoStat para gerar uma terminologia de óleo e gás por meio de *corpus*. Entretanto, o TermoStat só trabalha

com anotação gramatical. O segundo ponto a melhorar, destacado acima, envolve trabalhar com *corpus* anotado com dependência sintática também. Isso possibilitaria, por exemplo, a reproduzir o método de Hearst (1992), para identificar relações taxonômicas dentro do domínio.

A informação sintática foi utilizada em outros trabalhos relacionados, tais como o de Lopes et al. (2009) e Wendt et al. (2010). Quanto à extração automática de terminologia, para além de combinações gramaticais, a informação sintática pode ser usada para se extrair sintagmas candidatos a termos. Com a dependência sintática, é possível, por exemplo, extrair todos os sintagmas nominais de um *corpus*, o que não chegou a ser feito. Ainda sobre os trabalhos relacionados, também temos em vista fazer uma comparação entre o glossário de geologia gerado em Wendt et al. (2010) e a nossa taxonomia, para medirmos o que há em comum entre os domínios.

Por outro lado, uma validação de especialistas da área também se mostra necessária, não somente para avaliarmos a taxonomia, mas também para pensarmos em uma limpeza maior das entradas. Há ainda pontos mais específicos a melhorar, por exemplos: o encaixe de famílias pequenas dentro de outras maiores, a detecção de outros prefixos do tipo super, bio e micro, e variações de escrita da mesma palavra, por exemplo, quando ela vem com hífen e sem hífen, com acento ou sem acento, no feminino ou no masculino. Além disto, partimos de diversas fontes para estabelecer a taxonomia, logo, também seria interessante reconhecer quais itens, para além dos que vieram de terminologias – glossários e dicionários – são de fato termos do domínio e quais são itens relacionados.

Um aspecto relevante da pesquisa é que a taxonomia construída já não é tão somente uma taxonomia, isto é, ela não apenas apresenta relações hierárquicas, mas também relações de igualdade. Como o trabalho foi aprimorado com a inclusão de algumas relações de sinonímia, apresentamos assim, para não dizer que é uma leve ontologia, uma taxonomia enriquecida. Na medida em que as melhorias propostas aqui se concretizem, o trabalho pode ganhar uma estruturação ainda mais parecida com a de uma ontologia, isto é, ela pode acabar contendo diversos níveis hierárquicos e diversos tipos de informação, para além de hiperonímia.

Ainda vale lembrar da inclusão dos códigos da OpenWordNet-PT. Algumas entradas da taxonomia, possuem os sentidos da OpenWordNet-PT para que se diferenciem de outras entradas com a mesma mãe, como é o caso das duas entradas

para “processamento”, relatadas na seção 5.2.2.5. Com isso, a taxonomia, também entra em diálogo com um dos recursos linguísticos mais utilizados no mundo do PLN: a WordNet. Pretendemos continuar utilizando o recurso como auxílio na tarefa de desambiguação e espera-se que isso inspire outros trabalhos a consultarem a taxonomia.

Espera-se que, uma vez que recursos linguísticos possuem utilidade e relevância dentro do PLN, a dissertação possa servir como modelo para a construção de outros recursos linguísticos, tanto dentro da área do óleo e gás, como em outros domínios técnicos. Assim, apresentamos à comunidade do PLN, e ao ambiente corporativo da indústria de óleo e gás, a primeira versão de um recurso linguístico em português voltado para essa área, para os mais diversos fins.

7
Referências bibliográficas

ALLENDE-CID, Héctor. Machine learning: catalisador da ciência. **Revista da Sociedade Brasileira de Computação**, ed. 1, número 39, 2019.

ANTHONY, Laurence. AntConc: **Design and development of a freeware *corpus* analysis toolkit for the technical writing classroom**. Proceedings of Professional Communication Conference. 729-737. 10.1109/IPCC.2005.1494244. 2005.

BASSANI, Hansenclever F. O impacto da aprendizagem profunda na sociedade e academia. **Revista da Sociedade Brasileira de Computação**, ed. 1, número 39. 2019.

BICK, Eckhard. The Parsing System PALAVRAS: **Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework**. Aarhus University Press Aarhus. 2000.

BORDEA, Georgeta, LEFEVER, Els e BUITELAAR, Paul. **Semeval-2016 task 13: Taxonomy extraction evaluation** (texeval-2). In SemEval-2016, pages 1081–1091. Association for Computational Linguistics. 2016.

BORDEA, Georgeta, BUITELAAR, Paul, FARALLI, Stefano e NAVIGLI, Roberto. **Semeval-2015 task 17: Taxonomy extraction evaluation (TExEval)**. SemEval2015, 452(465):902. In Proceedings of the SemEval workshop, 2015.

CAMACHO-COLLADOS, José, DELLI BOVI, Claudio, ESPINOSA-ANKE, Luis, ORAMAS, Sergio, PASINI, Tommaso, SANTUS, Enrico, SHWARTZ, Vered, NAVIGLI, Roberto, SAGGION, Horacio. **SemEval-2018 Task 9: Hypernym Discovery**. Proceedings of The 12th International Workshop on Semantic Evaluation. 712-724. 10.18653/v1/S18-1115. 2018.

CLEVERLEY, P. H. & BURNETT, S. The best of both worlds: highlighting the synergies of combining manual and automatic knowledge organization methods to improve information search and discovery. **Knowledge Organization**, vol. 42, n. 6, 2015.

COHEN, K. Bretonnel, VERSPOOR, Karin, FORT, Karën, FUNK, Christopher, BADA, Michael, PALMER, Martha e HUNTER, Lawrence E. The Colorado Richly Annotated Full Text (CRAFT) Corpus: Multi-Model Annotation in the Biomedical Domain. In: Nancy Ide e James Pustejovsky (Ed.). **Handbook of Linguistic Annotation**. Dordrecht: Springer, 2017.

COHEN, K. Bretonnel, LANFRANCHI, Arrick, CHOI, Miji Joo-young, BADA, Michael, BAUMGARTNER, William A., PANTELEYEVA, Natalya, VERSPOOR, Karin, PALMER, Martha, e HUNTE, Lawrence E. Coreference annotation and resolution in the Colorado Richly Annotated Full Text (CRAFT) corpus of biomedical journal articles. **BMC Bioinformatics**, vol. 18, n. 372, pp. 1-14, 2017.

DODDINGTON, George, MITCHELL, Alexis, PRZYBOCKI, Mark, RAMSHAW, Lance, STRASSEL, Stephanie, WEISCHEDEL, Ralph. **"The automatic content extraction (ACE) program–tasks, data, and evaluation".** Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04). 2004.

DROUIN, Patrick. Term extraction using non-technical *corpora* as a point of leverage. In: **Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication**, vol. 9, n. 1, 2003.

FREITAS, Cláudia. "**Estudos linguísticos e Humanidades digitais: corpus e descorporificação**". Gragoatá, Niterói, v.22, n. 44, p. 1207-1227, 2017.

FREITAS, Cláudia, MOTA, Cristina, SANTOS, Diana, GONÇALO OLIVEIRA, Hugo e CARVALHO, Paula. **"Second HAREM: advancing the state of the art of named entity recognition in Portuguese".** LREC 2010, pp. 3630-3637. 2010.

FREITAS, Maria Cláudia de. **Elaboração automática de ontologias de domínio: discussão e resultados**. Tese (Doutorado em estudos da Linguagem), PUC-Rio, Rio de Janeiro, 2007.

GABOR, Kata, BUSCALDI, Davide, SCHUMANN, Anne-Kathrin, QASEMI ZADEH, Behrang, ZARGAYOUNA, Haïfa, CHARNOIS, Thierry. **SemEval-2018 Task 7: Semantic Relation Extraction and Classification in Scientific Papers.** Proceedings of The 12th International Workshop on Semantic Evaluation. 679-688. 10.18653/v1/S18-1111. 2018.

GOMES, Diogo S. M., CORDEIRO, Fábio e EVSUKOFF, Alexandre G. **Word embeddings em português para o domínio específico de óleo e gás**. Rio Oil & Gas Expo and Conference 2018. Rio de Janeiro, RJ, Instituto Brasileiro de Petróleo, Gás e Biocombustíveis (IBP), 2018.

GONÇALO OLIVEIRA, Hugo, SANTOS, Diana e GOMES, Paulo. Extracção de relações semânticas entre palavras a partir de um dicionário: o PAPEL e a sua avaliação. **Linguamática**. 2. 2010.

GRISHMAN, Ralph e SUNDHEIM, Beth. **Message Understanding Conference 6: A Brief History.** (in) *Proceedings of the 16th International Conference on Computational Linguistics*, *COLING 96*. Copenhaga, Dinamarca, p. 466-471, 1996.

GRISHMAN, Ralph e SUNDHEIM, Beth. **Design of the MUC-6 Evaluation**. (in) *Proceedings of the 6th Message Understandin, MUC-6.* Columbia, MD, EUA, pp. 413-422. 6-8 de Novembro de 1995.

HEARST, Marti & SCHUTZE, Hinrich. **Customizing a lexicon to better suit a computational task**. In ACL SIGLEX Workshop, Columbus, Ohio, 1993.

HEARST, Marti. **Automatic acquisition of hyponyms from large text corpora**. In: Proceedings of the 14th International Conference on Computational Linguistics, Nantes, 1992.

HOVY, E., MARCUS, M., PALMER, M., RAMSHAW, L., WEISCHEDEL, R.: **Ontonotes: the 90% solution**. In Proceedings of Human Language Technologies: The 2006 Annual Conference of the North American Chapter of the Association for Computational Linguistics, New York City: Companion Volume, Short papers, pp. 57–60, 2006.

IDE, Nancy & VÉRONIS, Jean. **Knowledge Extraction from Machine-Readable Dictionaries: An Evaluation.** In Petra Steffens, editor, Proceedings of Machine Translation and the Lexicon, Third International EAMT Workshop, Heidelberg, Germany. 19-34. 10.1007/3-540-59040-4_18. Springer: 1993.

JIANG, Ridong, BANCHS, Rafael E., LI, Haizhou, Abacha, BEN e ROTH, Dan. "**Evaluating and Combining Named Entity Recognition Systems**". In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1, pp. 694–702, 2016.

KILGARRIFF, A. **Thesauruses for Natural Language Processing**. Proceedings of NLP-KE, Beijing, China, 2003.

KIM, Jin-Dong, OHTA, Tomoko, TATEISI, Yuka, TSUJII, Jun'ichi. GENIA corpus: A semantically annotated corpus for bio-textmining. **Bioinformatics**. 19 (Suppl 1), i180-2. 10.1093/bioinformatics/btg1023. England: Oxford, 2003.

KRIEGER, Maria da Graça e FINATTO, Maria José. **Introdução à terminologia: teoria e prática.** São Paulo: Contexto, 2004.

Lafon, P. Sur la variabilité de la fréquence des formes dans un corpus, MOTS, no 1, pp. 128-165, 1980.

L'HOMME, Marie-Claude. **Lexical Semantics for Terminology: an introduction**. Terminology and Lexicography Research and Practice, volume 20. Amsterdam/Philadelphia: John Benjamins Publishing Company, 2020

LOPES, Lucelene, VIEIRA, Renata, FINATTO, Maria José, MARTINS, Daniel, ZANETTE, Adriano, RIBEIRO JR, Luiz Carlos**. Extração**

automática de termos compostos para construção de ontologias: um experimento na área da saúde. RECIIS – R. Eletrônica de Comunicação, Informação e Inovação em Saúde. Rio de Janeiro, vol. 3, n.1, p.76-88, 2009.

LOPES, Lucelene, OLIVEIRA, Leandro Henrique de & VIEIRA, Renata. **Portuguese term extraction methods: comparing linguistic and statistical approaches**. 2010.

LOPES, Lucelene, FERNANDES, Paulo, VIEIRA, Renata & FEDRIZZI, G. **ExATO lp - An automatic tool for term extraction from portuguese language *corpora***. Proc. of the 4th Language & Technology Conference (LTC '09). 427-431. 2009.

MANNING, Christopher D. Last Words: Computational Linguistics and Deep Learning. **Computational Linguistics**, vol. 41, n. 4, pp. 701-707, 2015.

MANNING, C. e SCHÜTZE, H. 1999. **Foundations of Statistical natural language processing.** Cambridge, MA: The MIT Press, 1999.

McENERY, Tony & HARDIE, Andrew. Corpus linguistics : method, theory and practice. Cambridge: Cambridge University Press, 2012.

MITKOV, Ruslan. **The Oxford handbook of computational linguistics**. Oxford: Oxford University Press, 2003.

MOTA, Cristina & SANTOS, Diana (eds.). **Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM.** Linguateca, 2008.

PACHECO, César Augusto Rodrigues & Natasha Sophie PEREIRA, . Deep Learning Conceitos e Utilização nas Diversas Áreas do Conhecimento. **Revista Ada Lovela**ce. V. 2, p. 34-49, 2018.

PARK, Y.; BIRD, R.; BOUGAREV, B. **Automatic Glossary Extraction: Beyond Terminology Identification**. Proceedingings of the 19th COLING, Taipei, Taiwan, 2002.

PRADHAN, S.; RAMSHAW, L.; MARCUS, M.; PALMER, M.; WEISCHEDEL, R.; Xue, N. "**Conll-2011 shared task: Modeling unrestricted coreference in ontonotes**". In: Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task, pp. 1–27, 2011.

ROBERTS, Angus, GAIZASUKAS, Robert, HEPPLE, M., GUO, Yikun. Combining **Terminology Resources and Statistical Methods for Entity Recognition: an Evaluation.** Proceedings of the Language Resources Evaluation Conference. Marrakech, Morocco, 2008.

SANTOS, Diana & Cardoso, Nuno (Eds) Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área. Linguateca. 2007.

SANTOS, Diana. "Corporizando algumas questões". In: Stella E. O. Tagnin & Oto Araújo Vale (eds.), **Avanços da Lingüística de Corpus no Brasil**. USP, São Paulo : Editora Humanitas, pp. 41-66, 2008.

SPÄRCK JONES, Karen “Computational linguistics: what about the linguistics?”, **Computational Linguistics**, Volume 33, n. 3, p.437-441, 2007.

SPÄRCK JONES, Karen. “Natural language processing: a historical review”. **Current Issues in Computational Linguistics: in Honour of Don Walker**, (Ed. A. Zampolli, N. Calzolari and M. Palmer), Amsterdam: Kluwer, pp. 3-16, 1994.

THOMPSON, Paul, ANANIADOU, Sophia e TSUJIITHE, Jun’ichi. **GENIA Corpus: Annotation Levels and Applications**. In: Nancy Ide e James Pustejovsky (Ed.). Handbook of Linguistic Annotation. Dordrecht: Springer, 2017.

TJONG KIM SANG, Erik F. **Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition**. In: Proceedings of CoNLL-2002, Taipei, Taiwan, pp. 155-158, 2002.

TJONG KIM SANG, Erik & DE MEULDER, F. 2003. **Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition**. In Proc. Conference on Natural Language. pp. 142-147. Edmonton, Canada, 2003.

VAN DER LAAN, Regina. **Tesauro e Terminologia: uma inter-relação lógica**. Tese (Doutorado). Universidade Federal do Rio Grande do Sul, Programa de Pós-Graduação em Letras. Porto Alegre, 2002.

WENDT, Igor, LOPES, Lucelene, MARTINS, Daniel, VIEIRA, Renata, LIM, Vera Lúcia Strube de. **Geração automática de glossários de termos específicos de um corpus de Geologia**. Ontobrás, 2010.

YADAV, Vikas & BETHARD, Steven “A Survey on Recent Advances in Named Entity Recognition from Deep Learning models”. COLING, 2018.

YAMADA, Ikuya, ASAI, Akari, SHINDO, Hiroyuki, TAKEDA, Hideaki e MATSUMOTO, Yuji. LUKE: **Deep Contextualized Entity Representations with Entity-aware Self-attention**. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020.

8
Anexos

ANEXO 1: Lista de modificadores utilizados como filtro de limpeza

realizado: teste realizado
utilizado: processo utilizado
preparado: solução preparada
produzido: água produzida
obtido: solução obtida
proposto: mecanismo proposto
envolvido: mecanismo envolvido
variado: processos variados
necessário: pressão necessária
aleatório: processo aleatório
adotado: ponto adotado
crítico: ponto crítico
diluído: solução diluída
estabelecido: limite estabelecido
injetado: água injetada
formado: água formada
causado: degradação causada
presente: água presente
anterior: ensaio anterior
final: produto final
principal: processo principal
maior: tempo maior
menor: concentração menor
direta: relação direta
de problema: solução de problema
de são: cabo de são
de santa: cabo de santa
de incerteza: relação de incerteza
de interesse: ponto de interesse
de estudo: desenvolvimento de estudo
exato: solução exata
ultra: água ultra
único: camada única
ótimo: solução ótima
figura: água figura
tabela: processo tabela
brasileiro: ensaio brasileiro
específico: taxa específica
estratégico: ponto estratégico
vermelha: camada vermelha
central: ponto central
chave: ponto chave
alternativo: processo alternativo
brilhante: ponto brilhante
vizinho: ponto vizinho
redondo: cabo redondo
experimental: ensaio experimental
tipo: elemento tipo