CAMINHOS EM LINGUÍSTICA APLICADA
Universidade de Taubaté - UNITAU
ISSN 2176-8625

LIMPEZA SEMI-AUTOMÁTICA DE LISTAS DE CANDIDATOS A TERMOS: UM EXERCÍCIO EM TERMINOLOGIA

Jackson Wilke da Cruz SOUZA

Universidade Federal de Alfenas – UNIFAL-MG

Resumo: O desenvolvimento de ferramentas e aplicações linguístico-computacionais por meio do Processamento Automático de Línguas Naturais (PLN) propiciou avanços eficientes e de destaque para os estudos nas áreas de Terminologia e Linguística de *Corpus*, permitindo, sobretudo, a análise de extensos *corpora* especializados e a extração de candidatos a termos considerando a combinação de n-gramas. Entretanto, o resultado das listas de candidatos a termos geradas apenas com base em conhecimento estatístico/probabilístico apresentam candidatos que não possuem características linguísticas para serem validados como termos de áreas especializadas do conhecimento. Isso deve-se ao fato de a maioria das abordagens utilizadas considerarem a frequência com que os candidatos ocorrem no *corpus* e, consequentemente, não refletem os padrões morfológicos de formação de palavras do Português do Brasil (PB). Assim, neste trabalho desenvolveu-se um algoritmo automático para a limpeza de listas de candidatos a termos que considere os padrões morfológicos do PB, levando em consideração a abordagem de autômatos finitos. Para tanto, utilizaram-se listas de n-grama (uni, bi, tri, tetra e pentagrama) de candidatos a termos a partir de um *corpus* sobre Revisão de textos. Como resultado, o algoritmo proposto possibilitou diminuir mais de 7 mil candidatos a termos das listas originadas com abordagem quantitativa. Tal resultado pode impactar positivamente trabalho desempenhado pelos terminológos, diminuindo o tempo de análise, encaminhando aos especialistas de domínio listas de candidatos a termo substancialmente menores e com melhores características linguísticas.
Palavras-Chave: Terminologia. Candidatos a termo. Limpeza de listas. PLN.

SEMI-AUTOMATIC CLEANING OF TERM CANDIDATE LISTS: AN EXERCISE IN TERMINOLOGY

Abstract: The development of linguistic-computational tools and applications through the Automatic Processing of Natural Languages (PNL) allowed efficient and outstanding advances for studies in the areas of Terminology and *Corpus* Linguistics, above all, the analysis of extensive specialized *corpora* and the extraction of term candidates considering the combination of n-grams. However, the result of lists of candidates for terms generated only based on statistical/probabilistic knowledge presents candidates who do not have linguistic characteristics to be considered terms of specialized areas of knowledge. This occur because most of the approaches used consider the frequency with which the candidates occur in the *corpus* and, consequently, do not reflect the morphological patterns of word formation in Brazilian

Portuguese (BP). Thus, in this paper, we developed an automatic algorithm for cleaning candidate lists of terms that considers the morphological patterns of BP, considering the Finite Automata approach. For that, we used n-gram lists (uni, bi, tri, tetra and pentagram) of candidates for terms from a *corpus* on Texts Review. As a result, the proposed algorithm made it possible to decrease more than 7 thousand candidates for terms from the lists originated with a quantitative approach. Such a result can positively impact the work performed by terminologists, reducing the analysis time, sending to domain specialists some lists of substantially smaller term candidates with better linguistic characteristics.
Keywords: Terminology. Term candidates. Cleaning lists. NLP.

LIMPIEZA SEMIAUTOMÁTICA DE LISTAS DE CANDIDATOS A PLAZO: UN EJERCICIO EN TERMINOLOGÍA

Resumen: El desarrollo de herramientas y aplicaciones lingüísticas-computacionales a través del Procesamiento Automático de Lenguas Naturales (PLN) proporcionó avances eficientes para los estudios en las áreas de Terminología y Lingüística del *Corpus*, permitiendo, sobre todo, el análisis de amplios cuerpos especializados y extracción de candidatos a término considerando la combinación de n-gramos. Sin embargo, el resultado de las listas de candidatos para los términos generados solo en base al conocimiento estadístico / probabilístico presenta a los candidatos que no tienen características lingüísticas para ser considerados términos de áreas especializadas de conocimiento. Esto se debe al hecho de que la mayoría de los enfoques utilizados consideran la frecuencia con la que los candidatos ocurren en el *corpus* y, en consecuencia, no reflejan los patrones morfológicos de formación de palabras en portugués brasileño (PB). Por lo tanto, en este trabajo se desarrolló un algoritmo automático para limpiar las listas de términos de candidatos que considera los patrones morfológicos de PB, teniendo en cuenta el enfoque de autómatas finitos. Para eso, utilizamos listas de n-gramas (uni, bi, tri, tetra y pentagrama) de candidatos para los términos de un *corpus* sobre Revisión de textos. Como resultado, el algoritmo propuesto hizo posible disminuir aproximadamente 57% de candidatos para los términos de las listas originadas con un enfoque cuantitativo. Tal resultado puede impactar positivamente el trabajo realizado por los terminólogos, reduciendo el tiempo de análisis, enviando listas de especialistas de dominio de candidatos de términos sustancialmente más pequeños con mejores características lingüísticas.
Palabras-clave: Terminología Término candidatos. Limpieza de listas. PLN

INTRODUÇÃO

Os avanços dos estudos em Terminologia e Linguística de Corpus têm permitido que paradigmas teórico-metodológicos sejam revistos e aprimorados, frente a ferramentas e recursos automáticos desenvolvidos por pesquisas em subáreas do Processamento Automático de Línguas Naturais (PLN). Nesse contexto, tais ferramentas podem ser capazes de (semi)automatizar, dinamizar, otimizar e, até mesmo, aprimorar os resultados de análises e

descrições terminológicas, já que utilizam majoritariamente abordagens estatísticas, o que permite analisar grandes volumes de dados linguísticos.

Atualmente, a criação de vocabulários e produtos terminológicos, em especial, baseiam-se em descrições sistemáticas de domínios específicos do conhecimento. Por conta de decisões de projeto, há produtos que não utilizam nenhuma abordagem computacional em seu processo de desenvolvimento. Essa decisão resulta no fato de o pesquisador ter de empenhar grande parte do tempo de descrição e observação dos dados em detrimento do menor alcance de descrição de dados, bem como a possibilidade de não analisar/considerar algum termo ou comportamento linguístico por ter ocorrido poucas vezes no corpus de domínio, já que o pesquisador terá de priorizar os casos/fenômenos que são mais recorrentes.

Dessa maneira, o papel dos recursos em PLN, em Terminologia, é auxiliar o terminólogo a desenvolver descrições linguísticas de maneira sistêmica (ou horizontalizada) e otimizada (ou verticalizada), de maneira que permita o pesquisador (i) desenvolver observações relevantes e simultâneas com base em um grande volume de textos (p.ex. Anthony (2014)), (ii) analisar corpora paralelos (p.ex. Santos (2008) e Teixeira et al. (2009)), (iii) criar de produtos terminológicos (p.ex. Almeida e Oliveira (2012)), ou mesmo (iv) consultar bases ontológicas (p.ex. Maziero et al. (2008)), dentre outras possibilidades.

Uma das tarefas que mais pode demandar tempo do pesquisador ao elaborar um produto terminológico (como dicionários e glossários) é a limpeza de listas de candidatos a termos. Recorrentemente, as listas de candidatos são geradas com base em cálculos estatísticos de frequência de n-grama. Assim, se a palavra “a” se repete frequentemente ao longo do corpus, ao gerar uma lista de unigrama, por exemplo, esse será um dos candidatos a termo. Um outro exemplo são possíveis candidatos a termos que não possuem estrutura morfológica adequada, mas, por conta da abordagem estatística, podem ocorrer nas listas, como “tubo de” e “arco da”, como exemplos de bigramas.

Além disso, na maioria das vezes o pesquisador pode não conhecer tecnicamente o domínio de conhecimento a que se propôs a produzir o produto terminológico. Assim, para limpar as listas de candidatos, deixando aqueles que indicam potencial a serem validados como termos pelos especialistas de domínio, o pesquisador terá de se basear em pistas linguísticas (como relações semânticas ou estruturas de palavras) nessa tarefa. Entretanto, como as listas de

ISSN 2176-8625

candidatos a termo geradas são extensas, comumente fazem-se cortes estatísticos para analisar as palavras mais frequentes, já que quanto menor o n-grama, a lista de candidatos tende a ser maior.

Como uma possível solução a esse tipo de problema, propõe-se, neste artigo, utilizar os pressupostos metodológicos de Bates (1978) para construção de um Autômato Finito (AF) que sirva de filtro linguístico, eliminando sequências de n-grama que possuam estruturas morfológicas inaceitáveis na estrutura do PB para que seja realizada a limpeza das listas de candidatos a termos. AF é uma abordagem que analisa objetos estado a estado, de acordo com condições pré-estabelecidas, como exemplificado na Figura 1.

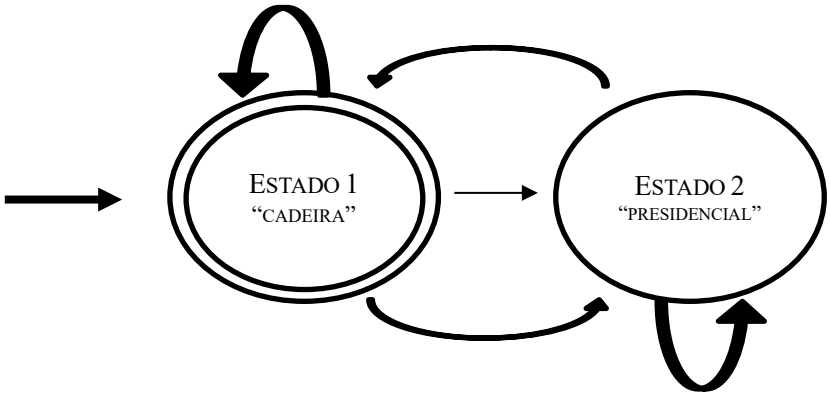

Figura 1 – Demonstração de Autômato Finito.
Fonte: Elaborado pelo autor.

Com base na Figura 1, tem-se duas possíveis palavras, a saber "cadeira" e "cadeira presidencial". Poder-se-ia utilizar o AF dessa figura para analisar a ocorrência das referidas palavras em um corpus, por exemplo. Assim, a análise começa a partir do estado inicial (nesse caso, Estado 1), que já é uma palavra válida. Outra alternativa é combinar o Estado 1 com o Estado 2, resultado em "cadeira presidencial", que também é uma palavra válida. Entretanto, ainda com base no exemplo, o AF pode propor a análise de "cadeira cadeira", por exemplo, já que sobre o Estado 1 há uma função de transição em looping, o que também ocorre com o Estado 2. Uma outra alternativa seria combinar o Estado 1 e apenas o Estado 2 em looping, resultando em "cadeira presidencial presidencial", por exemplo.

Ao aplicar essa abordagem na limpeza de candidatos a termos, objetiva-se aplicar o conceito de recursividade linguística às listas de candidatos, resultando na diminuição de ruídos

provocados durante o processo de extração de candidatos, além do tempo demasiado de análise que o terminológo empenha nessa tarefa. Para tanto, com auxílio da ferramenta linguístico-computacional Unitex, desenvolveu-se um AF que se baseia em traços e/ou padrões morfológicos do PB, que desempenhou a função de filtro linguístico para limpeza semiautomática de listas de n-grama de candidatos a termos. A fim de testar o AF construído, geraram-se listas de n-grama de candidatos a termos (uni, bi, tri, tetra e pentagrama) de um corpus sobre revisão de textos (BALESTERO, 2019), resultando em uma diminuição de mais de 7 mil candidatos.

Este artigo está organizado em 5 seções. Na segunda seção apresenta-se a relação do PLN e da Terminologia, ao fomentar ferramentas e recursos que auxiliem nas tarefas terminológicas. A terceira seção é responsável por delimitar a arquitetura para geração de produtos terminológicos, e a apresentação do ambiente E-Termos. A quarta seção apresenta a Teoria dos Autômatos, a qual pôde ser desenvolvida e aplicada por meio do software Unitex. Na quinta seção, apresentam-se teste e avaliação do recurso aqui desenvolvido. Por fim, na sexta seção, tecem-se algumas considerações finais.

1. PLN, TERMINOLOGIA E LINGUÍSTICA DE *CORPUS*: CONTRIBUIÇÕES INTERDISCIPLINARES

De acordo com Dias-da-Silva et al. (2007), desde a introdução dos computadores digitais na década de 40 na sociedade e, posteriormente, de sua popularização, diversos campos do conhecimento científico têm utilizado as ferramentas e recursos impulsionados e desenvolvidos pelo uso desses sistemas inteligentes. Em Linguística, essa introdução garantiu o desenvolvimento e/ou aprimoramento de subáreas que sistematizaram e ampliaram os estudos descritivos da linguagem, como a Tradução Automática e a Sumarização Automática (produção automática de resumos), por exemplo, consolidando o Processamento Automático de Línguas Naturais (PLN).

Nesse contexto, a Terminologia beneficia-se com a produção de ferramentas, recursos e abordagens em PLN que visam propor soluções linguístico-computacionais capazes de lidar com os desafios e demandas de descrições linguísticas e elaboração de produtos terminológicos.

Sobre a delimitação da Terminologia, de acordo com Sager (1993), compreende-se o conjunto de práticas e métodos utilizados na compilação, descrição, gestão e apresentação dos

termos de uma dada modalidade especializada da linguagem, bem como áreas do conhecimento que estejam em advento. Ademais, segundo Lorente (2004), o objetivo da Terminologia seria compreender o funcionamento de unidades lexicais especializadas em determinadas contextos comunicativos profissionais, acadêmicos ou científicos.

Em retrospecto, os primeiros trabalhos terminológicos eram desenvolvidos substancialmente de maneira manual (p.ex. Almeida, 2000), consequência de não haver promoção e valorização dos estudos da linguagem em conjunto às teorias e metodologias computacionais, o que resultava em ferramentas computacionais específicas não estarem à disposição dos terminólogos para que tivessem seus trabalhos auxiliadas por elas. Dessa maneira, tarefas de coleta e manipulação de grandes volumes de textos (corpora), análise de candidatos a termos, elaboração e organização de mapas conceituais, análise de termos e protocolo de fichas terminológicas eram prejudicados em amplitude e profundidade, já que tais tarefas não dispunham de ferramentas computacionais bem adaptadas às tarefas desenvolvidas pelo terminólogo.

Um dos grandes avanços que o PLN permitiu e, consequentemente, a Terminologia auxiliou por conta de seus objetivos e motivações, foi a promoção dos estudos sobre Linguística de Corpus (LC). De acordo com Dipper (2008), a Linguística computacional e a LC desenvolveram-se quase que de maneira correlata. Segundo o autor, de modo geral, as áreas exploram corpora eletrônicos a fim de obter informações e dados linguísticos reais e em larga escala, variando apenas entre abordagens qualitativas ou quantitativas.

A partir de uma perspectiva quantitativa (ou empiricista), a LC oferece subsídio teórico-metodológico ao PLN, já que ao analisar um conjunto de textos em formato eletrônico, é possível extrair e descrever dados e fenômenos linguísticos relevantes. Assim, treinam-se algoritmos que mineram informações em corpora que sejam representativos da língua ou de uma variedade/modalidade linguística específica, resultando na aquisição/extração de informações que permitem a identificação de unidades terminológicas em seus respectivos contextos de uso. Por outro lado, sob uma abordagem qualitativa, os estudos da linguagem permitem descrever (ou mesmo prever) comportamentos linguísticos a partir de uma coleção de dados.

Deve-se ressaltar que, de acordo com Lüdeling e Kytö (2008), pesquisas baseadas em corpus não devem deixar de lado a introspecção linguística. Nesse caso, um conjunto de textos,

por exemplo, (i) dará suporte à pesquisa para a verificação de hipóteses levantadas, que podem, inclusive, ser refutadas; ou (ii) será o meio pelo qual o pesquisador irá explorar conhecimentos linguísticos, munindo-se de informações sobre usos de palavras, sintagmas ou construções frásticas de maneira qualitativa; ou ainda (iii) pode ser utilizado com o objetivo de extrair e/ou descrever informações extralinguísticas, as quais não estão explícitas na superfície textual, mas que podem influenciar na maneira com que o autor de um texto expressa-se, como o gênero textual ou fatores temporais.

No Brasil, em especial, sob a perspectiva da Teoria Comunicativa da Terminologia (TCT) (CABRÉ, 2005), a Terminologia passou a subsidiar o desenvolvimento de produtos com base em grandes corpora. Essa abordagem advém da ideia de que, no corpus, é possível observar fenômenos linguísticos em seus contextos comunicativos reais, garantindo, a princípio, naturalidade e veracidade linguística às informações extraídas.

Aluísio e Almeida (2006) apontam que a utilização de corpora em trabalhos terminológicos justifica-se pelo fato de (i) fontes estruturadas (como dicionários e glossários) serem bastante raros, a depender da variedade linguística a ser sistematizada (como domínios emergentes), e (ii) os corpora permitirem a extração de dados em variados contextos de ocorrência e em grandes quantidades, aprimorando suas definições.

Ao apontar diretrizes sobre o uso de corpus para a construção de ontologias, Di-Felippo e Souza (2010) indicam que os conjuntos de textos devem ser projetados em função da pesquisa a que eles servirão de suporte ou fonte de extração de conhecimento. Os autores definem que os textos devem pertencer a um recorte sincrônico da língua, além de serem relativos a um determinado domínio. Ademais, deve-se atentar a características de composição do corpus, como autenticidade, representatividade, diversidade e balanceamento (SARDINHA, 2004). Assim, consideram-se os postulados da própria LC e as decisões baseadas em características do domínio a ser descrito durante a produção de qualquer produto terminológico.

Como resultado da contribuição teórico-metodológica das áreas de PLN, Terminologia e LC, tem-se a semiautomatização das tarefas necessárias para gerenciamento e elaboração de produtos terminológicos, como dicionários e glossários. Assim, na próxima seção, apresenta-se uma arquitetura (genérica) para geração de produtos terminológicos a partir de corpora,

baseando-se na apresentação do ambiente on-line e colaborativo E-Termos, no qual implementa-se tal arquitetura.

2. PLN, TERMINOLOGIA E LINGUÍSTICA DE *CORPUS*: CONTRIBUIÇÕES INTERDISCIPLINARES

Como visto, as pesquisas terminológicas que resultam na elaboração de produtos representativos de uma porção real da língua pautam-se em estudos de *corpora*. Nessa perspectiva, Aluísio e Almeida (2006) propõem a metodologia que é brevemente descrita nesta seção, a qual é implementada por Almeida e Oliveira (2012) na plataforma *online* E-Termos, como meio material de contribuições interdisciplinares entre o PLN, Terminologia e LC.

Com o objetivo de tornar acessíveis todas as tarefas, o E-Termos (ALMEIDA; OLIVEIRA, 2012) tem como pressuposto tornar colaborativa, gratuita e semiautomática a elaboração de produtos terminológicos, já que, no ambiente, todas as tarefas compreendem abordagens da Terminologia, da LC e do PLN. Ademais, por ser colaborativo, é possível que as tarefas desenvolvidas sejam compartilhadas entre os pesquisadores cadastrados em cada projeto. Assim, o pesquisador responsável libera acessos diferentes a outros pesquisadores de seu grupo, permitindo, por exemplo, que alguns estejam responsáveis pela elaboração do *corpus*, enquanto outros se responsabilizam apenas pelas relações conceituais do mapa conceitual.

Dentre as sete tarefas propostas, ressalta-se que a quantidade delas que, de fato, serão utilizadas em um produto terminológico depende exclusivamente de decisões traçadas a cada projeto. Por exemplo, para a construção da ontologia TermiNet do domínio da Educação à Distância (DI-FELIPPO, 2010) foram adotadas as Tarefas 1 a 4, por conta da natureza e objetivo do projeto; já durante a construção do Dicionário Terminológico de Materiais Cerâmicos (ALMEIDA, 2000; ALMEIDA et al., 2011) foram adotadas todas as sete tarefas.

ISSN 2176-8625

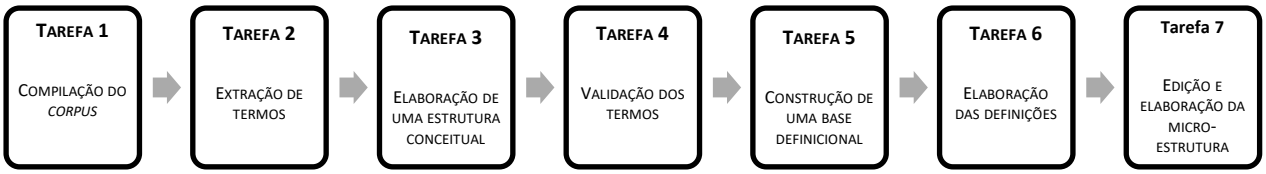

Figura 2 – Arquitetura genérica para geração de produtos terminológicos.
Fonte: Elaborado pelo autor.

A maior vantagem em utilizar o ambiente E-Termos para a elaboração de produtos terminológicos é o fato de o pesquisador não necessitar de conhecimentos prévios específicos sobre computação ou estatística, já que, na maioria das tarefas, estão implementadas ferramentas linguístico-computacionais para auxiliá-lo de maneira semiautomática.

2.1. COMPILAÇÃO E ANÁLISE DE QUALIDADE DO CORPUS

A proposta de compilação semiautomática é permitir o usuário utilizar ferramentas de busca de textos na Web em que, por meio de palavras-chave, recupere textos acerca dos temas investigados. Entretanto, essa funcionalidade ainda está sendo implementada no ambiente, sendo possível apenas fazer manualmente o *upload* de textos ou fragmentos deles para compor o *corpus*. Além disso, é possível analisar a qualidade/características do *corpus* (em termos de quantidade de *tokens* e/ou *types* e metadados de publicação, por texto, por exemplo), permitindo que o usuário faça adequações no conjunto de textos, compilando-os em um ´nico arquivo, ou excluindo aqueles que, eventualmente, não serão necessários (Tarefa 1, Figura 2).

2.2. EXTRAÇÃO AUTOMÁTICA DE CANDIDATOS A TERMOS

Nessa etapa, o pesquisador pode extrair candidatos a termos com base no *corpus* compilado, de maneira que, por meio de modelos estatísticos de extração de n-gramas, são geradas as listas. Nessa tarefa ainda pode-se editar (dentro e/ou fora do próprio ambiente) as listas, uma vez que elas podem apresentar ruídos por conta da abordagem estatística (Tarefa 2, Figura 2).

ISSN 2176-8625

2.3. EDIÇÃO DO MAPA CONCEITUAL E CATEGORIZAÇÃO DE TERMOS

Nessa etapa, o pesquisador pode criar e editar um mapa conceitual com base nas listas de n-gramas de termos, permitindo visualizar, editar e criar relações conceituais (como hiponímia ou meronímia) entre os candidatos a termos. O ambiente permite, nessa tarefa, criar um perfil de usuário a um especialista de domínio, o qual poderá avaliar e, eventualmente, editar as listas de termos e relações conceituais entre os candidatos (Tarefa 3, Figura 2).

2.4. GERENCIAMENTO DA BASE DE DADOS TERMINOLÓGICOS

Nessa tarefa é possível que o usuário crie campos que comporão a ficha terminológica de um termo (como *termo*, *contexto*, *modalidade* e *gênero textual*), além de possibilitar a elaboração de sua base definicional. Salienta-se que o pesquisador não tem a obrigatoriedade de construir tais recursos com base no mapa conceitual, eventualmente construído na etapa descrita em 3.3 (Tarefas 4 e 5, Figura 2).

2.5. INTERCÂMBIO E DIFUSÃO DE TERMOS

Por fim, nessa etapa o pesquisador poderá editar os verbetes, o que possibilitará a difusão, intercâmbio e consulta dos produtos terminológicos (Tarefas 6 e 7, Figura 2). Para tanto, o usuário conta com um conjunto de ferramentas que facilita o acesso ao *corpus* ou a fontes estruturadas acrescidas por ele mesmo para produzir o texto da definição proposta, além de ser possível escolher alguns aspectos visuais do dicionário ou glossário final (como cor e fonte) e quais termos estarão disponíveis aos consulentes na versão final.

2.6. ESPECIFICIDADES SOBRE A TAREFA DE EXTRAÇÃO AUTOMÁTICA DE CANDIDATOS A TERMO

Especificamente sobre a Etapa 3 (Figura 2), Almeida e Oliveira (2012) apontam que há três abordagens para a extração de candidatos a termos, a saber (i) abordagem estatística, (ii) abordagem linguística e (iii) abordagem híbrida.

Em (i), o modelo utilizado se baseia em cálculos estatísticos para o reconhecimento de combinações possíveis de n-gramas. Assim, quando o usuário solicita a criação de uma lista de bigrama com base em um *corpus* do domínio da Fisioterapia, por exemplo, é possível que tenha

como resultado "*articulação de*", "*bolsa com*" e "*a cápsula*". Tais "termos" passam, a partir desse procedimento, a compor uma primeira e possível versão da lista de candidatos a termos.

Em (ii), o modelo tem por prioridade aplicar padrões linguisticamente recorrentes para a identificação de unidades terminológicas, como "*nome + adjetivo*". Assim, ainda no cenário hipotético da Fisioterapia, em uma lista de bigrama, em que se aplica o padrão linguístico exemplificado, seriam recuperadas as combinações "*líquido sinovial*" ou "*membrana fibrosa*", por exemplo. Salienta-se que, neste caso, seria necessário um *corpus* anotado morfossintaticamente ou uma ferramenta computacional que pré-processasse os textos com tais informações linguísticas.

Por fim, em (iii), o modelo prevê a combinação entre as abordagens de (i) e (ii). Dessa maneira, tem-se a criação de um sistema que detecta estruturas linguísticas e, após a identificação, aplicam-se conhecimentos estatísticos para a decisão da relevância ao domínio. Assim, os bigramas "*líquido sinovial*" e "*membrana fibrosa*" somente comporiam a possível lista de candidatos a termos após terem sua relevância estatística comprovada (ou seja, ser recorrente no *corpus*), aliada ao padrão linguístico "*nome + adjetivo*".

Todas as três abordagens apontadas têm impactos diretos nos métodos e modelos utilizados por sistemas em PLN, podendo ser classificados em abordagens profunda (quando o sistema possui muito conhecimento linguístico) ou superficial (quando não há muito conhecimento linguístico empregado). As aplicações que possuem melhores desempenhos nessas abordagens, no entanto, são as que dependem de uma descrição exaustiva de informações linguísticas. Para tanto, Almeida e Vale (2010) indicam que o sistema linguístico deve ser submetido a constantes descrições antes e durante a extração dos candidatos a termos, já que essa tarefa é um dos critérios para validação dos candidatos.

No E-Termos, a extração de candidatos a termos é puramente estatística, já que a quantidade de candidatos extraídos depende diretamente do tamanho do *corpus*. Nessa perspectiva, Lopes et al. (2009) propõe uma razão aritmética entre a quantidade de *tokens* e a frequência com que ocorrem no *corpus*. Com base nesse cálculo, os autores indicam que as unidades extraídas do conjunto de texto com menor frequência não conferem valor terminológico ao candidato. Tal proposta baseia-se em Estopá-Bagot (1999), em que se defende

que em alguns domínios especializados as unidades terminológicas ocorrem com frequências altas, justificando eliminar da análise candidatos com frequência muito baixa.

Entretanto, por não possuir uma abordagem profunda quanto à composição da lista de candidatos a termos, o método de Lopes et al. (2009) produz listas de n-grama com ruídos, como os casos exemplificados anteriormente, na abordagem puramente estatística. Isso resulta na composição de uma lista em que os candidatos não possuem características morfológicas para serem validados como termos, como é o caso de "*bolsa com*".

Assim, propôs-se neste artigo, como medida a suprir essa lacuna, a criação de um AF que auxilie a limpeza das listas de n-grama de candidatos a termos, resultando em versões menores e com menos ruídos provocados durante a elaboração com abordagens estatísticas. Assim, na próxima seção, delimitam-se os aspectos teóricos sobre os AF e suas aplicações computacionais em estudos da linguagem.

3. AUTÔMATOS FINITOS PARA APLICAÇÕES LINGUÍSTICAS

De acordo com Ranchhod (2001), as demandas impostas ao PLN exigem que os pesquisadores produzam descrições linguísticas sistemáticas e completas, de acordo com seus respectivos fins. Isso, segundo a autora, visa diminuir possíveis falhas durante o processamento automático (como a geração de listas de candidatos a termos) advindas da insuficiência de dados linguísticos (como a falta de informação morfossintática atrelada aos *tokens* do *corpus*) em etapas anteriores do projeto. Assim, para iniciar um trabalho/projeto linguístico-computacional ou utilizando alguma ferramenta dessa natureza para determinada aplicação, a autora sugere que o léxico e sua devida caracterização/classificação são componentes importantes a qualquer sistema de PLN.

Ranchhod (2001) ainda aponta que o aprimoramento de descrições linguísticas em suas aplicações específicas será possível a partir de construções de dicionários e gramáticas[1] que atendam a especificidade da aplicação. Dessa maneira, é preciso identificar as unidades lexicais (como palavras que estão unidas por hífen ou palavras que estejam separadas por espaços em

[1] Aqui, *dicionário* é tido como o repositório eletrônico de unidades lexicais com suas devidas informações linguísticas (morfossintáticas, por exemplo); enquanto *gramática* é o repositório eletrônico de regras de organizações e/ou combinações entre as unidades lexicais de uma língua.

branco), descrever as informações linguísticas atreladas às unidades e resolver ambiguidades provenientes da homografia, e, como fruto desse processo, ter-se um dicionário.

Entretanto, tendo como pressuposto a recursividade linguística em termos de criatividade e produtividade (CHOMSKY, 1956), é inviável que, a cada análise linguística, sejam desenvolvidos novos dicionários e/ou gramáticas para processar um *corpus*. Nesse sentido, os AFs são formalizações computacionais capazes de preservar o princípio de recursividade linguística, proporcionando o "reaproveitamento" acerca do conhecimento em pesquisas posteriores.

Ao realizar análises sintáticas, Bates (1978) utiliza a *Augumented Transition Netwok* (ATN), formalismo proposto inicialmente por Woods (1970). Assim, dado uma sentença, os modelos linguísticos desenvolvidos em ATN visam realizar uma análise sintática em que se observa a organização dos componentes necessários a uma sentença, em que em uma entrada (via escrita), os modelos investigam padrões sintáticos entre os componentes, com o objetivo de descrever uma gramática local[2] para a sentença sob análise.

Compreende-se uma gramática construída em ATN como um conjunto de estados finitos, em que cada estado representa uma categoria linguística específica (como categorias sintagmáticas, por exemplo), a qual ocorre na superfície textual da sentença.

(1) O homem comprou um carro.

Sob uma perspectiva sintática, (1) é composto por um Sintagma Nominal (no caso "*O homem*") e um Sintagma Verbal (no caso, "*comprou um carro*"). Tais sintagmas são compostos por constituintes que, no caso do nominal seria "o" e "homem", enquanto no verbal seria o verbo e seu complemento. Na árvore sintática, cada nó representa um estado de análise do AF (Figura 3).

[2] O termo *gramática local* é utilizado recorrentemente em trabalhos que se baseiam na construção de AFs. Nesse sentido, constroem-se regras específicas à descrição linguística que está sendo realizada, não estando sujeita necessariamente às regras propostas pela Gramática Tradicional, por exemplo.

ISSN 2176-8625

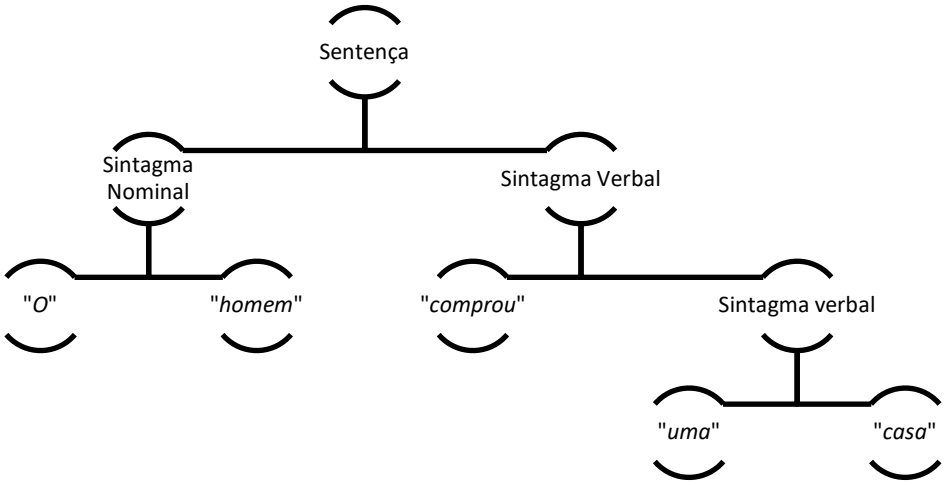

Figura 3 – Exemplo de análise sintática em esquema arbóreo.
Fonte: Elaborado pelo autor.

Com base nessa concepção, Bates (1978) determina que a *Recursive Transition Network* (RTN) se estabelece a partir da junção das gramáticas baseadas em ATN, uma vez que, do ponto de vista computacional, a recursividade é construída ao utilizar fragmentos dos estados finitos de uma gramática local para realizar outras análises sintáticas.

(2) O senhor comprou um avião.

Em (2), tem-se os Sintagma Nominais na posição agentiva (“O senhor”) e passiva (“um avião”), os quais são distintos daqueles utilizados nas mesmas posições em (1), mantendo-se o verbo “comprar”. Assim, ao construir uma gramática local para o exemplo em (2), é possível realizar uma análise sintática com base na gramática já existente em (1), acrescentando, apenas, os novos estados finitos e não previstos nela, como exemplificado na Figura 4.

ISSN 2176-8625

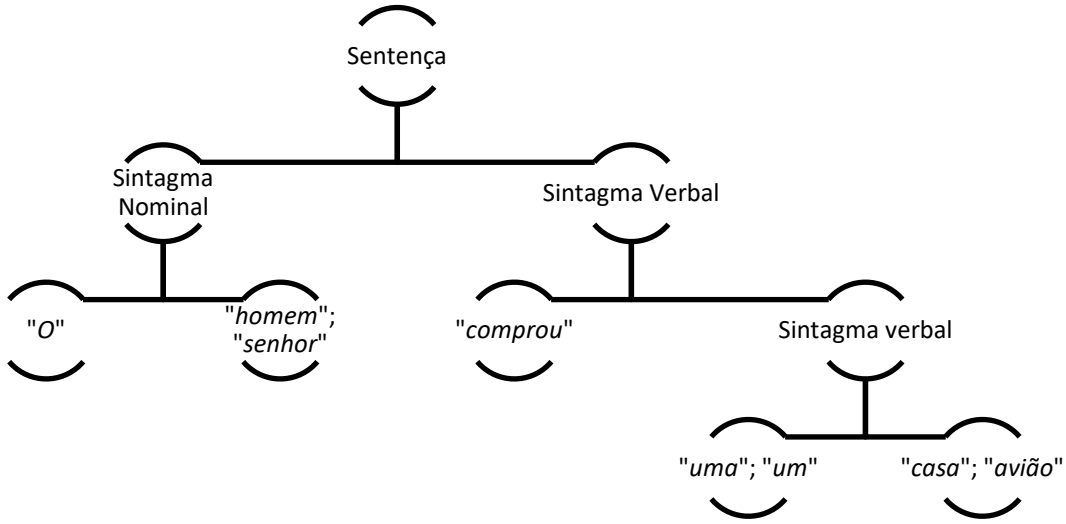

Figura 4 – Exemplo de análise sintática recursiva em esquema arbóreo.
Fonte: Elaborado pelo autor.

Dessa maneira, com essa abordagem, é possível implementar e representar em sistemas computacionais conhecimentos de níveis linguísticos, como a análise sintática. Ainda do ponto de vista computacional, os modelos de análise que se baseiam em AF podem ser remodelados/redesenhados em função de outras aplicações linguísticas, apenas tendo a necessidade de acrescentar novos elementos às regras gramaticais, como demonstrado na Figura 4. Nesse sentido, não haveria a necessidade de iniciar um novo AF a cada novo projeto e/ou aplicação, bastando apenas acrescentar novos elementos às regras já existentes, ou mesmo ampliando-as com base naquelas que já constam no autômato.

Na próxima subseção apresenta-se a ferramenta Unitex/GramLab 3.1, a qual possibilitou a construção semiautomática do AF utilizados neste artigo.

3.1. CARACTERÍSTICAS (LINGUÍSTICAS) DO AF

Para a construção do AF, utilizou-se a ferramenta linguístico-computacional Unitex/GramLab 3.1[3] (PAUMIER, 2002). Esse *software* foi desenvolvido com o objetivo de processar *corpora* linguísticos por meio da disponibilização de dicionários eletrônicos ao usuário, e permitir que ele construa e edite dicionários e gramáticas locais.

[3] Disponível em: http://unitexgramlab.org/pt. Acesso em: 27/02/2020.

Com relação a outras ferramentas automática de processamento de *corpus*, o Unitex destaca-se pelo fato de já ter implementado a ele um dicionário com informações morfossintáticas. Assim, dado um novo texto, a ferramenta o pré-processa, reconhecendo as informações morfossintáticas relacionadas a todas as entradas, as quais já estão previstas no dicionário da ferramenta. Caso as entradas não constem, o usuário poderá inseri-las no dicionário local da ferramenta e, a partir de então, o Unitex, instalado na máquina local, passará a considerá-la/reconhecê-la.

Pelo fato de o usuário conseguir editar o dicionário da ferramenta, é possível também que ele construa o seu próprio com informações que lhe parecer pertinentes, além da Morfossintaxe. Assim, é possível que o usuário crie um léxico com informações semânticas e processar/analisar o *corpus* considerando Papeis Temáticos, por exemplo.

Com base no dicionário *default* do Unitex/GramLab 3.1, é possível realizar buscas entre as combinações de "*Verbo + Objeto direto"* e analisar tal consulta usando o concordanciador[4] da ferramenta. Entretanto, caso o usuário disponha de um *corpus* anotado a nível de *constituintes sintáticos*, será necessário construir uma gramática específica para investigar os fenômenos linguísticos a partir dessa anotação, por exemplo. Dessa maneira, além da edição e criação de dicionários, é possível desenvolver gramáticas locais no *software*.

Dado isso, para construir AFs que atuem de maneira semiautomática na tarefa de limpeza de listas de candidatos a termos, é necessário assumir dois aspectos, a saber: (i) processos de formação de palavras em uma língua específica e (ii) especificações linguísticas da área de conhecimento que está sendo estudada. É importante ressaltar esses dois aspectos já que (i) prevê processos de formações que não são aceitáveis pela língua estudada e (ii) abrange casos específicos do domínio pesquisado. Tal reflexão é importante pois, ao estabelecer que se deve retirar da lista de candidatos seguimentos de caracteres terminados com –*y*, por exemplo, deve-se estudar a relevância e o impacto desse filtro no domínio específico, já que é possível que se esteja excluindo da análise algum candidato importante que, posteriormente, poderia compor a lista de termos daquele domínio.

[4] Concordanciador é uma ferramenta disponível na maioria dos processadores automáticos de *corpora*, a qual permite, após a busca de uma palavra ou a combinação entre palavras, visualizar a ocorrência regular da busca. Isso permite, por exemplo, ver o ambiente textual em que ocorre a sequência "*líquido sinovial*", dada a quantidade de caracteres à esquerda e/ou à direita do termo de busca.

ISSN 2176-8625

Dessa forma, o AF desenvolvido neste artigo foi construído de maneira genérica, considerando apenas os processos de formação de palavras do PB, ressaltando que é possível fazer as adaptações possíveis ao domínio em que ele poderá ser aplicado, como demonstrado na Figura 5.

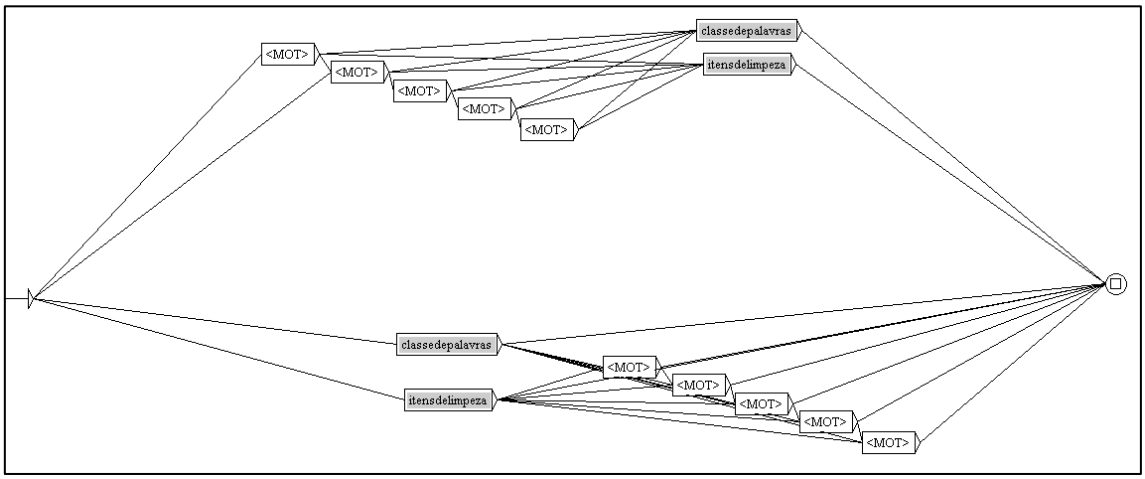

Figura 5 – Arquitetura do Autômato Finito para limpeza de candidatos a termos.
Fonte: Elaborado pelo autor.

A partir da arquitetura proposta para o AF (Figura 5), tem-se a aplicação da recursividade linguística, uma vez que é possível combinar caminhos entre os elementos dispostos no autômato. Assim, é possível utilizar selecionar uma sequência qualquer de carácter (ou "MOT") seguido por uma classe de palavras específica, como "preposição" e eliminá-la da lista de candidatos a termo, resultando na análise de um possível bigrama. Para que não seja elaborado uma nova opção de verificação que percorra um caminho semelhante, pode-se acrescentar à análise anterior mais uma sequência de caracteres, resultando na observação de trigrama (no caso, *MOT + MOT + preposição)*. Além disso, é possível eliminar alguns itens das listas utilizando a recursividade já apontada. Por meio da combinação entre n-grama e caracteres que podem ter se corrompido durante o processo de conversão do documento em *.txt* (como "*", "º", e "$") ou mesmo letras e/ou sílabas que não iniciam palavras em português (como "k", "y", "w" e "ç"), foi possível criar opções de análise com o AF para eliminação dessas sequências das listas de candidatos.

Por fim, ainda sobre a Figura 5, pode-se utilizar apenas uma única vez os grafos "ClasseDePalavras" e "ItensDeLimpeza". Entretanto, isso não foi demonstrado apenas por questões visuais/gráficas de representação neste artigo.

ISSN 2176-8625

3.2. APLICAÇÃO E CONFIGURAÇÃO DO AF

Para utilizar o AF desenvolvido, será preciso a instalação prévia do *software* Unitex/GramLab 3.1[5]. Após isso, são necessárias três etapas, a saber: (i) pré-processamento dos textos, (ii) *upload* e utilização do AF e (iii) limpeza das listas.

a) Pré-processamento das listas de candidatos a termo

Uma vez que o Unitex trabalha com textos eletrônicos (formato *plain text*), é necessário fazer com que *software* passe a considerar a lista de candidatos a termos como um texto a ser pré-processado. Nos textos, é comum ter delimitadores que indicam o fim das sentenças (como *ponto final*); ao verificar isso, o Unitex insere automaticamente, logo após o delimitador, a marca de "{S}". Já que as listas de candidatos a termo não possuem sentenças e, consequentemente, também não tem pontuação, será necessário inserir manual e previamente o delimitador "{S}" para que processe a lista considerando-a o *corpus* de análise.

b) Upload e utilização de AF e lista de candidatos a termos

Após o pré-processamento manual, as listas de candidatos a termos, ainda sujas, deverão ser inseridas no diretório "*Corpus*", na máquina local, para que *software* entenda-as como o conjunto de textos a ser analisado. Em seguida, deverá ser copiado o AF "*LimpezaDeCandidatoTermo*[6]" para o diretório "*Processing*" e, assim, na interface gráfica do programa, será necessário fazer o *upload* da lista de candidatos a termos e o arquivo de limpeza da lista.

c) Limpeza das listas de candidatos a termo

O AF desenvolvido neste trabalho servirá de filtro linguístico, retirando tudo aquilo que for recuperado pelo autômato. Assim, no campo de busca, na interface gráfica do Unitex, deverá ser utilizado o arquivo "*LimpezaDeCandidatoTermo*", seguindo a sequência de abas *Text* > *Located Sequences* > *Extract unmatching units*, e escolher um arquivo em branco *.txt*, para que

[5] Para mais detalhes sobre instalação e utilização da ferramenta, consultar o manual disponível em: http://www.usp.br/gmhp/soft/unitx.pdf
[6] Disponibiliza-se o AF no seguinte endereço: https://github.com/jackcruzsouza/LimpezaCandidatosTermo

sejam impressas somente as ocorrências que não foram recuperadas pelo autômato; ou seja, a lista limpa de candidatos a termo.

4. TESTE E AVALIAÇÃO

A fim de avaliar o AF construído e a abordagem utilizada, testou-se a aplicação do autômato em listas de candidatos a termos extraídos a partir de um *corpus* de domínio. Para tanto, escolheu-se o *corpus* sobre revisão de textos (BALESTERO, 2019), o qual é composto por textos técnicos em PB. De acordo com a autora, a abordagem para a criação de um vocabulário para esta área é baseada na TCT, e os candidatos a termos (uni, bi, tri, tetra e pentagrama) foram extraídos automaticamente no ambiente E-Termos. Na Tabela 1, apresentam-se as listas de candidatos com suas respectivas extensões.

Tabela 1 – Relação entre a Quantidade de Candidatos a termo e N-Grama no conjunto de teste.

N-GRAMA	QNT. DE CANDIDATOS A TERMO	
	ANTES	DEPOIS
UNIGRAMA	5812	1535
BIGRAMA	710	378
TRIGRAMA	2477	201
TETRAGRAMA	476	20
PENTAGRAMA	121	10

Fonte: Elaborado pelo autor.

Na Tabela 1, demonstra-se a quantidade de candidatos em função das combinações de n-grama. Como apresentado na subseção 3.2, para a elaboração das listas não se consideram informações linguísticas, mas somente o número de ocorrências de dada combinação no *corpus*. Dessa maneira, na lista de bigrama, por exemplo, ocorrem as combinações de sequência de caracteres "*à + adequação*" ou "*à + análise*" por serem recorrentes no *corpus*, as quais, evidentemente, não se configuram como termos da área de Revisão de Textos, tampouco colaboram para a compreensão do domínio pelo terminológo.

Após a aplicação do AF, há redução de mais de 7 mil "candidatos" cobertos pelo autômato (de 5812 candidatos a 2144). A eliminação automática de palavras que não possuem chance a serem termos garante que algumas palavras sejam mais bem evidenciadas nas listas de candidatos, como é o caso de "ABNT", na lista de unigrama: antes posicionada entre as trinta palavras mais recorrentes da lista, passa a ocupar a nona linha do arquivo.

Entretanto, apesar da redução bastante significativa entre as listas das Tabelas 1 e 2, ressalta-se que o trabalho de verificação das listas de candidatos pelo terminológo não é dispensável, mas somente aprimorado e otimizado. Há combinações que ainda terão características linguísticas específicas do domínio, e que poderão ser reconhecidas e/ou excluídas somente com o trabalho conjunto entre terminológo e especialista de domínio. Portanto, a utilização desse AF caracteriza-se como uma abordagem semiautomática, uma vez que as listas finais de termos sejam ainda mais reduzidas ao serem submetidas às avaliações manuais.

CONSIDERAÇÕES FINAIS

O objetivo deste trabalho foi desenvolver uma aplicação com abordagem computacional que pudesse dinamizar e otimizar o trabalho do terminológo durante o processo de limpeza de listas de candidatos a termos geradas automaticamente. Ressalta-se que o AF aqui proposto deverá sofrer adaptações levando em conta a área de conhecimento em que será aplicado, incrementando lhe informações linguísticas (como a eliminação de determinados tipos de verbos, bem como suas flexões e conjugações). Entretanto, esse tipo de aprimoramento só poderá ser realizado após as primeiras utilizações do AF nas listas de candidatos a termos de determinado domínio e após ser realizados estudos preliminares acerca do comportamento linguístico do campo em questão.

Além disso, a cobertura dos casos alcançados pelo presente AF poderá ser ampliada devido os dicionários em PB do Unitex poderem ser atualizados, inserindo-se novas entradas, em especial de domínios emergente, como o da Revisão de Textos. Tal apontamento é percebido na diminuição percentual entre as listas de unigrama e pentagrama de candidatos a termo das Tabelas 1 e 2, por exemplo: por não considerar particularidades linguísticas da área de conhecimento em questão, observou-se que quanto maior é a combinação de n-grama, menor foi a diminuição de candidatos (apenas de 8,26% para pentagrama), o que é inversamente factível considerando tamanhos menores de combinação de n-grama (26,41% para unigrama). Entretanto, ressalta-se que a diminuição das listas foi significativa, mesmo baseando-se em aspectos linguísticos generalistas.

Como trabalho futuro, objetiva-se promover modificações no AF aqui desenvolvido, visando aumentar a capacidade de limpeza de listas por meio do enriquecimento de padrões morfossintáticos, bem como a ampliação de entradas no dicionário do *software* utilizado.

AGRADECIMENTOS

Agradeço a colaboração e as discussões teórico-metodológicas dos professores Gladis Maria de Barcellos Almeida e Oto Araújo Vale, e das colegas Camila Azevedo e Mirella Balestero, da Universidade Federal de São Carlos.

REFERÊNCIAS

ALMEIDA, G. M. B.; LANDIM-KAMIKAWACHI, D. S.; MANFRIN, A. M. P.; SOUZA, I. P.; IZUMIDA, F. H.; DI-FELIPPO, A.; ZAUBERAS, R. T.; MELCHÍADES, F. G.; BOSCHI, A. O. Glossário de revestimento cerâmico. In: I. M. ALVES (org.). **Cadernos de Terminologia**. 1ª. ed. São Paulo: FFLCH-USP, v.4, p.03-56, 2011

ALMEIDA, G.M.B**. Terminologia Comunicativa**: uma aplicação com vistas à elaboração de um glossário de Materiais Cerâmicos. Tese de Doutorado. Araraquara: UNESO, Faculdade de Ciências e Letras, 2000.

ALMEIDA, G.M.B.; OLIVEIRA, L.H.M. Terminology and computational linguistics: new praxes in terminography. **Cahiers de Lexicologie**, Paris, v. 101, p. 139-153, 2012.

ALMEIDA, G.M.B.; VALE, O.A. Do texto ao termo: interação entre Terminologia, Morfologia e Linguística de *Corpus* na extração semiautomática de termos. In: ISQUERDO, A.N.; FINATTO, M.J.B. (Orgs.). **As ciências do léxico**: lexicologia, lexicografia, terminologia. 2ª. ed., v. IV, p. 483-499. Editora da UFMS e editora da UFRGS, Campo Grande/Porto Alegre. 2010.

ALUISIO, S.M.; ALMEIDA, G.M.B. O que é e como se constrói um *corpus*? Lições aprendidas na compilação de vários *corpora* para pesquisa linguística. **Calidoscópio**, v. 4, n. 3, p. 156-178, 2006.

ANTHONY, L. **AntConc**. Computer software. Tokyo/Japão: Waseda University, 2014.

BALESTERO, M.S. **Definições terminológicas da Revisão de Textos**: estudos iniciais para a elaboração de um glossário. Dissertação de Mestrado. Araraquara: UNESP, Faculdade de Ciências e Letras, 2019.

BATES, M. The theory and pratice of augmented transition networks. In BOLC, L. (ed.) Natural Language Comunication with computers. LNCS, vol.63, p. 191-252. Springer, Heidelberg, 1978.

CABRÉ, M. T. A Terminologia, uma disciplina em evolução: passado, presente e alguns elementos de futuro. **Debate Terminológico**, n. 01, 2005.

CHOMSKY, N. Three models for the description of language. **IRE Transactions on information theory**, v. 2, n. 3, p. 113-124, 1956.

DIAS-DA-SILVA, B.C.; MONTILHA, G.; RINO, L. H. M.; SPECIA, L.; NUNES, M. D. G. V.;DE OLIVEIRA JR, O. N.; PARDO, T.A.S. Introdução ao Processamento das Línguas Naturais e Algumas Aplicações. **Série de Relatórios do Núcleo Interinstitucional de Linguística Computacional**, v. 3, 121 p., 2007.

DI-FELIPPO, A. The TermiNet Project: an Overview. In **Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas**. Los Angeles/California, p. 92–99, 2010.

DI-FILIPPO, A.; SOUZA, J.W.C. O projeto do *corpus* para a construção de uma *wordnet* terminológica. In: Shepherd, T.; Sardinha, T.B.; Veirano-Pinto, M. (Org.). **Caminhos da Linguística de *Corpus***. Campinas: Mercado de Letras, v. 01, p. 225-245, 2010.

DIPPER, S. Theory-driven and *corpus* driven computational linguistics, and the use of *corpora*. In LUDELING, A.; KYT, M. (edit.). ***Corpus* Linguistics**: An international handbook, v.1, p.68–96. Mouton de Gruyter/Berlin. 2008

ESTOPÁ-BAGOT, R. **Extracció de terminologia: elements per a construcció d'un SEACUSE** (Sistema d'Extracció Automàtica de Candidats a Unitats de Significació Especialitzada). Tese de Doutorado – Universidade Pompeu Fabra, Barcelona, 1999.

LOPES, L.; FERNANDES, P.; VIEIRA, R.; FEDRIZZI, G. ExATOlp: an automatic tool for term extraction from Portuguese language *corpora*. In: **Proceedings of International Language and Technology Conference**, Poznam/Poland, p. 1-5, 2009.

LORENTE, M. A lexicologia como ponto de encontro entre a gramática e a semântica. In: ISQUERDO, A.N.; KRIEGER, M.G (orgs). **As ciências do léxico**, vol. II. Campo Grande: Editora UFMS, 2004.

LÜDELING, A.; KYTÖ, M. ***Corpus* linguistics**. Walter de Gruyter, 2008.

MAZIERO, E. G.; PARDO, T.A.S. A.; DI-FELIPPO. DIAS-DA-SILVA, B. C. A base de dados lexical e a interface web do TeP 2.0: thesaurus eletrônico para o Português do Brasil. In: **Proceedings of the XIV Brazilian Symposium on Multimedia and the Web**. ACM, p. 390-392. Vila Velha/Brasil, 2008.

PAUMIER, S. Unitex user manual, 2002. Disponível em https://unitexgramlab.org/releases/2.1/man/Unitex-GramLab-2.1-usermanual-en.pdf. Acesso em: 27/02/2020.

RANCHHOD, E.M. O uso de dicionários e de autómatos finitos na representação lexical das línguas naturais. In **Tratamento das Línguas por Computador**: Uma Introdução à Linguística Computacional e suas Aplicações, p. 13-47, 2001.

SAGER, J.C. **Curso prático sobre el procesamiento de la terminología**. Madrid: Fundación Germán Sánchez Ruipérez/Pirámide, 1993.

SANTOS, D. **Corpos linguísticos da Linguateca**: apresentação, 2008. Disponível em: https://www.linguateca.pt/documentos/SantosWorkshopTaLC2008.pdf. Acesso em: 27/02/2020.

SARDINHA, T.B. **Linguística de *corpus***. Manole: São Paulo, 2004.

TEIXEIRA, E.D.; SANTOS, D.; TAGNIN, S.E.O. CorTrad: um novo *corpus* paralelo multiversão para o par de línguas português-inglês. In **Proceedings of VIII Encontro de Linguística de *Corpus***. Rio de Janeiro, 2009.

WOODS, W. A. Transition network grammars for natural language analysis. In **Communications of the ACM**, v. 13, n. 10, p. 591-606, 1970.

Jackson Wilke da Cruz SOUZA

Doutor em Linguística (2019) pelo Programa de Pós-Graduação (PPGL) da Universidade Federal de São Carlos (UFSCar), na linha de pesquisa Descrição, Análise e Processamento Automático de Línguas Naturais (PLN). Mestre (2015) e Bacharel (2013) em Linguística pela UFSCar, com ênfase em PLN, atuando principalmente nas subáreas de Linguística de *corpus*, Semântica computacional, Terminologia, Sumarização automática e Análise textual. Atualmente é professor adjunto na Universidade Federal de Alfenas (UNIFAL-MG), campus avançado de Varginha.

Recebido em 02/04/2020 - Aceito em 31/05/2020