Processamento de Linguagem Natural, Linguística de *Corpus* e Estudos Linguísticos: uma parceria bem-sucedida

Natural Language Processing, *Corpus* Linguistics and Linguistics: a successful partnership

Maria José Bocorny Finatto[*]
Lucelene Lopes[**]
Alena Ciulla[***]

RESUMO: Neste artigo, apresentamos um exemplo de pesquisa que integra Processamento de Linguagem Natural (PLN) e Estudos Linguísticos, demonstrando que essa é uma associação possível e benéfica. Utilizamos uma ferramenta para extração de informações relevantes e representação de conteúdo a partir de *corpora* em português, o ExATOlp. Nessa iniciativa, foi utilizado como *corpus* o texto em português do *Curso de Linguística Geral*, de Saussure, para a investigação dos principais termos relacionados a conceitos importantes contidos nessa obra.

PALAVRAS-CHAVE: Processamento de Linguagem Natural. Linguística de *Corpus*. Estudos Linguísticos.

ABSTRACT: This article presents a research, in which is exemplified the integration between Natural Language Processing (NLP) and Linguistic Studies of the Text, with the aim of demonstrating that an association is possible and beneficial. A special tool, the ExATOlp, is used to extract relevant information and content representation from *corpora* in Portuguese. In that initiative it was used as a *corpus* the text in Portuguese of the *Course in General Linguistics*, in order to investigate the main terms related to important saussurean concepts in that book.

KEYWORDS: Natural Language Processing. *Corpus* Linguistics. Linguistics.

1. Introdução

Nosso objetivo, neste artigo, é demonstrar, mediante um exemplo pontual de pesquisa em andamento, que integrar o Processamento de Linguagem Natural (PLN) com os Estudos Linguísticos, via Linguística de *Corpus*, é um processo que pode ser muito proveitoso para todos os envolvidos, especialmente para os linguistas abertos ao contato com o novo. Entretanto, é preciso dizer que essa integração ou cooperação nem sempre tende a ser bem aceita, sendo recebida com restrições de ambas as partes. Afinal, trata-se de um diálogo entre cientistas da Computação e cientistas da Linguística ainda pouco disseminado no Brasil. Esses cientistas em contato representam, respectivamente, um diálogo entre as assim chamadas Ciências Exatas e as Humanidades.

[*] Docente do PPG-Letras-UFRGS e pesquisadora CNPq.
[**] Professora colaboradora da FACIN-PUCRS e pós-doutoranda DOCFIX-FAPERGS/CAPES.
[***] Professora visitante do Departamento de Linguística, Filologia e Teoria Literária e do PPG- Letras-UFRGS e pós-doutoranda DOCFIX-FAPERGS/CAPES.

A distância de escopos, acompanhada da diferença de pontos específicos de estudo e de epistemologias, tende a tornar-se uma barreira para concretizar alguma reciprocidade, sobretudo em um cenário universitário como o do Brasil, no qual os estudos linguísticos parecem ter pouca visibilidade para a população em geral, incluindo professores do ensino médio e fundamental, para além do que seja identificado, pelo leigo, como "estudar Gramática". Além disso, a interdisciplinaridade Humanas/Exatas ainda parece um ideal, sem contar que determinados tipos de pesquisas, como as que associam as áreas de Ciências Exatas e Biomédicas, por exemplo, parecem receber mais fomento do que outras parcerias.

Alguns cientistas da Computação, conforme percebemos em nossas experiências de interação, tendem a considerar o diálogo com o linguista algo bastante penoso. Isso porque percebem que o foco de ação do linguista parece geralmente incidir sobre uma problematização, sendo pouco centrado em modos de solução. Alguns linguistas, por sua vez, quando cooperam com cientistas de PLN, tendem a apresentar críticas bastante intensas quanto a uma, assim dita, ingenuidade (linguística) embutida em uma série de produtos e de sistemas computacionais que lidam com a linguagem humana. Assim, conforme vemos, propicia-se a não integração.

Martins (2011), ao tratar dessa inter-relação de saberes, sendo linguista com formação em PLN, elenca uma série de dificuldades enfrentadas pela pesquisa em Linguística Computacional, também conhecida como PLN[1]. Entre essas dificuldades, afirma que os resultados dos aplicativos e ferramentas linguístico-computacionais "ainda que possam ser extremamente úteis, especialmente quando envolvidas habilidades linguísticas mecânicas e repetitivas, longe estão de poder ser considerados verdadeiramente adequados." Do mesmo modo, Dias-da-Silva (1998), sendo matemático e cientista da Computação, com formação posterior em Letras e Linguística, já relatava toda uma série de dificuldades que o pesquisador de PLN tenderia a enfrentar quando buscasse a cooperação por parte de um linguista.

Ao tentar advogar pela validade da conciliação, tomaremos como leitor deste artigo um linguista com formação restrita às Letras e à Linguística, não muito familiarizado com o PLN nem com metodologias estatísticas e computacionais para a descrição da linguagem. Assim, para começar, importa situar o escopo e as especificidades do PLN.

[1] Há uma diferença sutil entre PLN e Linguística Computacional, mas não é o caso de aprofundá-la aqui.

Em primeiro lugar, é importante salientar que o processamento automático de línguas naturais, também denominado *Processamento de Língua/Linguagem Natural* ou *Linguística Computacional* (nomes reunidos aqui sob a abreviatura PLN), denota especificamente o objeto da pesquisa de desenvolvimento de sistemas computacionais capazes de processar objetos de natureza linguística (DIAS-DA-SILVA, 1996). Em segundo lugar, julgamos necessário que o nosso leitor compreenda que Linguística de *Corpus* não é sinônimo de PLN, visto que é a referência mais próxima que tem o linguista para um trabalho extensivo e com apoio de aplicativos computacionais.

A Linguística de *Corpus* (LC) é uma Linguística, como qualquer outra, situada no âmbito que conhecemos como Linguística Aplicada. Seu diferencial é o estudo da língua em uso, verificado o uso sempre em grande escala, com apoio informatizado e tratamento estatístico. Como um tipo de Linguística, tem o objetivo de ajudar os linguistas a entender como funciona a língua. Seu principal objetivo é a descrição de usos. Um ponto muito importante a frisar é que os linguistas de *corpus* partem de uma concepção de língua, que é entendida como um sistema probabilístico de combinatórias (para mais detalhes ver BERBER SARDINHA, 2004). Além disso, a padronização e a colocabilidade em torno de uma dada palavra são elementos centrais para a sua descrição em meio ao sistema da língua.

Assim, as ferramentas e métodos computacionais da LC estão a serviço de descrever, em larga escala, os diferentes usos que perfazem a linguagem, revelando respostas, mas também colocando muitas novas perguntas (e problemas) para o investigador que se debruce sobre a investigação de algum fenômeno linguístico. Nesse sentido, a Linguística de *Corpus* situa-se no âmbito da Linguística Aplicada, podendo ser entendida tanto como uma abordagem, quanto como uma metodologia. Suas implicações teóricas e metateóricas são de longo alcance, conforme Rajogopalan (2007). Ao privilegiar a observação estatística em larga escala e a identificação de padrões de uso e de combinatórias entre palavras a partir de extensas coleções de amostras de usos de línguas, a LC lança mão dos *corpora* ou *corpus*. Objetos centrais, definidores dessa Linguística, esses são os grandes acervos de textos em formato digital reunidos criteriosamente para um dado fim de estudo, de modo que representem a língua em suas diferentes possibilidades de acontecimento.

Por outro lado, o PLN mostra-se como uma área de investigação em Ciências da Computação, situando-se como uma subárea da Inteligência Artificial. Conforme Rosa (2011), o PLN pode ser definido como a habilidade de um computador em processar a

mesma linguagem que os humanos usam no seu dia-a-dia. Embora também lide com *corpus*, seus interesses são essencialmente aplicados. Para o PLN, não é objetivo descrever e, sim, criar soluções para problemas bastante pontuais, relacionados com o reconhecimento e a reprodução da linguagem humana em alguma escala. As soluções do PLN são sempre pensadas em termos de menor custo e maior benefício. Além disso, soluções diferentes devem ser comparadas entre si, avaliadas em termos de precisão, abrangência e margens de erro embutidas. Tais margens de erro, próprias dos enfoques estatísticos e lógicos, são sempre esperadas, especialmente quando ferramentas computacionais simulam desempenhos humanos, tais como a tradução, a elaboração de resumos, a avaliação de erros, a simulação e o reconhecimento de fala.

Outro argumento importante em favor de estudos que associem PLN e Linguística fundamenta-se no fato de que, de acordo com Branco et al. (2012), a pluralidade da linguagem é um dos maiores patrimônios da humanidade, mas, diante do mundo globalizado, representa uma das maiores barreiras na comunicação, tanto no dia a dia, quanto na esfera dos negócios ou da política e, acrescentamos, da educação e da ciência. Pela necessidade de transpor essa dificuldade, relatam os autores, a União Européia gasta cerca de um bilhão de euros por ano em traduções e interpretações de textos e na mineração de dados. "A tecnologia da linguagem e a investigação sobre as línguas naturais podem dar um contributo decisivo para se ultrapassarem essas barreiras linguísticas" (BRANCO ET AL., 2012, p.1) dizem os autores, que enxergam exatamente nessa conjunção que vimos descrevendo, entre as ciências da Computação e a Linguística, o futuro e o potencial de derrubada desse bloqueio de acesso ao conhecimento.

Para tanto, é preciso, em primeiro lugar, analisar sistematicamente as particularidades linguísticas das diferentes línguas, para que as ferramentas computacionais de apoio - inclusive os tradutores automáticos, mas não apenas - possam ser desenvolvidas, especializadas e adaptadas. É com o intuito de fazer esse levantamento, no que diz respeito à língua portuguesa, que Branco et al. (2012) se dedicam nessa publicação. Ressaltamos, aqui, apenas algumas das informações no que diz respeito à importância da língua portuguesa no mundo: além de ser uma das 23 línguas oficiais da União Europeia, é uma língua falada por cerca de 220 milhões de pessoas, é língua administrativa e de trabalho de 27 organizações internacionais e é a quinta mais utilizada na internet. No entanto, o português ainda é um idioma em que se dispõe de pouco ferramental computacional especializado, o que faz com que a maioria dos sistemas

computacionais não tenha um desempenho tão bom para o português, o que reforça, também nesse ponto, o aspecto positivo da pesquisa nessa área.

2. PLN e Estudos Linguísticos: uma associação possível e útil

Como mencionamos na introdução deste artigo, não só é possível, como também pode ser muito útil alguma cooperação entre colegas linguistas e cientistas de PLN. É preciso, no entanto, ter em mente, de maneira clara, os objetivos, limitações e benefícios de cada um dos parceiros de integração.

Retomando a discussão de Martins (2011), que, a nosso ver, traz uma visão um tanto pessimista a esse respeito, entendemos que já temos mudanças no cenário. Não é nosso objetivo, aqui, traçar o estado da arte da Inteligência Artificial (IA), tampouco revisar exaustivamente as contribuições do PLN, mas mostrar o quanto é possível e adequado o diálogo.

Nos primórdios da IA, criou-se a expectativa de que fosse possível fazer do computador uma máquina que conseguisse raciocinar e tomar decisões, como numa metáfora do cérebro humano. Isso porque, alavancados em grande parte pelo sucesso da máquina de Turing, os estudiosos em IA mostraram que os computadores não eram objetos que realizavam apenas cálculos aritméticos, mas aparentemente podiam simular comportamentos humanos. Um dos conceitos mais recentemente desenvolvidos em IA e que apresentou bom desempenho é o de *conexionismo*, que apresenta um modelo matemático simplificado do que seria o cérebro humano, permitindo a realização de várias tarefas em conexão, simultaneamente.

No entanto, independentemente das questões filosóficas sobre a possibilidade de comparação entre homem e máquina, e ainda que tenhamos muitos avanços no que diz respeito à complexidade de realização de tarefas simultâneas e à rapidez dos programas de computador, o problema parece residir no fato de que a máquina funciona apenas para um estado de coisas predefinido. Tal problema se aplica também e de maneira crucial ao tratamento computacional da linguagem humana, quando, sabemos, essa é uma atividade que envolve mudanças, associações, improvisos e conceitos que são constantemente reformulados pela interação entre os falantes e que não podem ser estabelecidos *a priori*. Tal impasse, ao contrário de ser visto como uma anomalia, pode ser visto como um desafio. Além disso, esse é o ponto principal da contribuição da Linguística para o PLN,

qual seja, o de fornecer dados linguísticos que a máquina não é capaz de inferir, mas pode, em parte, processar, melhorando o seu desempenho.

E, justamente por ter se deparado com esse impasse, grande parte das vertentes da IA deixou de lado a tentativa de desenvolver máquinas "inteligentes", voltando-se mais para o desenvolvimento de ferramentas de auxílio ao ser humano em análises e tomada de decisões. Uma das áreas que mais tem se destacado nesse sentido é o PLN, com aplicativos que, conforme Branco et al. (2012), "vieram ajudar ainda mais a automatizar e a facilitar o processamento da linguagem e comunicação", entre os quais destacamos os editores de texto e seus diversos recursos e aplicações, que vão desde o uso pessoal até o uso em larga escala pelas editoras, gráficas e escritórios de tradução, as mensagens eletrônicas em *e-mails*, *chats* e redes sociais, as ferramentas de busca por palavras-chave, o *data mining* (mineração de dados) e as ferramentas de tradução automática.

Ou seja, se de um lado houve, um dia, a expectativa de que a máquina reproduzisse o comportamento humano ao pensar, tomar decisões e se comunicar por meio da linguagem, essa não é a tarefa a que se propõe o PLN e muito menos a Linguística. Assim, ilustrando a caracterização do que se propõe ou não a ser uma ferramenta de PLN, vale dizer que o *Google Translator*, por exemplo, não foi criado para traduzir poemas, mas, sim, para ajudar uma pessoa a entender um texto de jornal em uma língua sobre a qual esse falante tem pouco ou nenhum conhecimento. Parece-nos, então, que os colegas mais pessimistas ou reticentes em relação à integração entre essas ciências estabelecem uma expectativa dentro de uma versão forte da IA, enquanto nós tendemos a perceber e a interagir com pesquisadores de PLN que têm consciência das limitações da máquina e das diferentes tarefas e contribuições que a Computação e a Linguística têm ao tratar a linguagem.

Martins (2011) indica que PLN e Linguística devem ser vistos como disciplinas autônomas, com métodos e objetivos próprios, com o que também estamos de acordo. Discordamos, porém, quando o autor diz que, sob a ótica do PLN, deveria ser investigada a possibilidade de replicar o dinamismo e a instabilidade próprios da linguagem. Nota-se aí um retorno à visão forte de IA, que alimentava a ideia de replicação de um cérebro humano, que é o único "aparelho", que se conhece até hoje, capaz de produzir e interpretar o dinamismo e a instabilidade próprios da linguagem. Ao que nos parece, o autor não fornece informações que indiquem como se poderia chegar a esse modelo ideal. Em nossa opinião, em prol de um diálogo que é cada vez mais demandado, é interessante que nos concentremos no que pode ser feito e traz resultados práticos - mesmo que não perfeitos.

Martins (2011), assim como vários outros estudiosos, duvida, então, "de um diálogo profícuo e interdisciplinar entre Linguística e Inteligência Artificial". Em nosso trabalho, pensamos o oposto. Partimos do princípio de que a tecnologia da linguagem, de acordo com Branco et al. (2012), é uma tecnologia facilitadora. O objetivo do estudioso em PLN é o de criar ferramentas robustas que deem conta de uma série de tarefas que envolvem processamento de linguagem natural, especialmente as multilíngues e/ou as que implicam grandes volumes de texto e informação. Enquanto linguistas, nesse trabalho associado ao PLN, valemo-nos de informações recuperadas pelos sistemas automáticos para nossas análises e descrições da língua, especialmente em se tratando de um grande volume de dados, e contribuímos para a melhoria desses sistemas, com a consideração de regras linguísticas que possam ser a eles integradas. Essas informações, por serem geradas automaticamente, com base em formalizações diferentes, estão longe de espelhar um conjunto perfeito de dados para o linguista. Por outro lado, por melhor que seja a compreensão de um fenômeno linguístico, raramente ele poderá ser formalizado em toda a sua complexidade em linguagem de máquina, ou linguagem de programação. Por isso, os parâmetros que usamos para julgar o desempenho de um tradutor automático, por exemplo, não são os mesmos de que nos valemos ao avaliar o desempenho de um tradutor humano. Contudo, não há como negar a importância desse tipo de recurso no âmbito do que ele se propõe, que é auxiliar o ser humano em tarefas que envolvam a linguagem natural, especialmente no que diz respeito ao tempo de processamento.

Feita essa contextualização, passamos agora à apresentação de um projeto de pesquisa por nós desenvolvido.

3. PLN e Linguística: um exemplo de integração

3.1 Contextualização do projeto de pesquisa

Embora ainda em curso e iniciado há pouco, nosso trabalho que associa PLN, Linguística de *Corpus* (LC) e Estudos Linguísticos pode servir aqui como exemplo, pois teve seu mérito reconhecido, como ideia, em um edital de fomento a pesquisas que envolvessem interdisciplinaridade Humanas/Exatas e alguma inovação tecnológica. Foi o único contemplado na área de Letras/Linguística na UFRGS no ano de 2012.

O projeto intitulado *Recuperação da informação em representação do conhecimento em bases de textos científicos de Linguística e de Medicina* iniciou em novembro de 2012 e foi contemplado por uma bolsa para pós-doutorado DOCFIX,

subsidiada pela CAPES e pela FAPERGS. Nessa investigação interdisciplinar, associam-se Letras/Linguística e Ciência da Computação/Processamento da Linguagem Natural. São explorados dois *corpora* de textos científicos em português: um de Medicina, na subárea das Pneumopatias Ocupacionais, e outro de Linguística, que é o texto em português do *Curso de Linguística Geral* (CLG) de F. de Saussure, organizado por C. Bally e A. Sechehaye. Ambos os *corpora* estão sendo tratados linguística e computacionalmente com vistas à representação automática do seu conteúdo e à sistematização de sua informação terminológica e textual.

A escolha desses dois *corpora* em especial - um de Medicina, outro de Linguística - foi guiada pela hipótese principal de que há diferenças entre o tratamento de textos científicos de áreas médicas e de ciências humanas, como a Linguística, de modo que se pretende detectar diferenças, formalizá-las e colocá-las em um sistema automático para a representação de conteúdo sob a forma de ontologias. As diferenças a formalizar alcançam as características dos diferentes gêneros discursivos envolvidos no estudo, considerando que o material de Medicina inclui artigos, teses, dissertações, textos de popularização para leigos e legislação relacionados a Pneumopatias Ocupacionais.

Com esse estudo, buscamos estabelecer parâmetros para subsidiar programas computacionais, tendo em vista um melhor desempenho em diferentes frentes de investigação que lidam com a linguagem científica escrita (ensino, descrição linguística, representação do conhecimento). Além da comparação entre os *corpora* de Medicina e de Linguística, pretendemos estudar cada *corpus* em separado, com diferentes objetivos. No que diz respeito ao material de Pneumopatias Ocupacionais, por ser ele composto de uma miscelânea de gêneros discursivos, investigaremos mais a fundo a questão da tipificação por gêneros discursivos. Acreditamos que os parâmetros que encontrarmos para a classificação de gêneros podem também auxiliar na automatização da extração e organização de informações dos textos - especialmente nesse caso, que se trata de Saúde Pública e as informações precisam ser divulgadas da maneira mais ampla, rápida e eficiente possível.

Neste trabalho, limitamo-nos ao trabalho com o *corpus* do CLG. Afora se tratar de obra fundadora da Linguística Moderna e um texto de complexidade e profundidade singulares, também particular é a concepção deste livro, pois que foi escrito com base em notas de alunos e por autores que não o próprio Saussure. E, somando-se a essa múltipla interpretação de que foi resultado o CLG, trabalhamos com uma tradução de tal texto, de que, é sabido, decorrem novas interpretações e releituras. Muitos dos conceitos tratados

no CLG foram, ainda, motivo de controvérsia e discussão, como em Culler (1979), Bouquet (1997), Normand (2000), Jäguer (2003) e Trabant (2005), entre outros, especialmente a partir da descoberta dos manuscritos de Saussure. A tradução do CLG para o português brasileiro foi feita no final da década de 60 e não foi feita uma revisão sistemática dessa tradução desde então.

Muitos assuntos podem, então, ser objeto de pesquisa, em se tratando do CLG. De nossa parte, iniciamos pela investigação desse *corpus*, que é o enfoque deste artigo, para um reconhecimento inicial da obra como um todo. Prosseguimos com o estudo sobre a referência aos principais conceitos-chave em Saussure, considerando um trabalho de cotejo entre a tradução brasileira que conforma o *corpus* e o seu original em francês. Observamos, no entanto, que não é nossa intenção propor uma nova tradução, mas apenas apontar alguns aspectos de tradução que possam ter influenciado a recepção da leitura tradutória no que diz respeito às ideias de Saussure no Brasil. Alguns desses problemas já foram por nós apontados em Ciulla; Finatto (2013) e fazem parte de uma das frentes de pesquisa dentro do projeto.

3.2 O trabalho com o CLG: aproximação da obra como *corpus* de estudo

A pesquisa apresentada neste artigo integra mais proveitosamente os recursos do PLN com os estudos linguísticos: trata-se da investigação dos principais termos relacionados a conceitos importantes no CLG. Partimos do pressuposto de que um panorama em larga escala, como o que as ferramentas automáticas permitem, pode proporcionar uma outra visão do CLG, que, até hoje, vem sendo analisado em detalhes e em partes segmentadas. Por isso, quanto à metodologia, lançamos mão, de um lado, do ferramental e do tratamento automático da LC, que já são excelentes para o grande volume de dados com que contamos. No entanto, eles não são sofisticados o suficiente para um tratamento refinado dos *corpora* do ponto de vista estatístico, conforme já mencionado. Tampouco podemos produzir, com ferramentas como geradores de contextos ou listadores de palavras e de *chunks*[2], algum recurso que permita uma visualização mais abrangente do conteúdo desses *corpora*. Assim, nos valemos da LC para os parâmetros de frequência e combinatória, por exemplo, mas avançamos rumo ao PLN, elegendo uma ferramenta em especial: o ExATOlp - *Extrator Automático de*

[2] Chunks, em PLN, são sequências de duas ou mais palavras que operam como uma unidade. Essas unidades não são processadas palavra por palavra, mas são armazenadas e puxadas da memória como um conjunto.

Termos para Ontologias em Língua Portuguesa. Trata-se de uma ferramenta computacional que é ao mesmo tempo aplicável a qualquer domínio, dirigido a textos escritos em português e implementa diversas técnicas avançadas de PLN que foram propostas no contexto da tese de doutorado de Lopes (2012), desenvolvida no grupo de PLN da PUCRS, liderado pelas professoras e pesquisadoras Renata Vieira e Vera Strube de Lima. Informações sobre o grupo podem ser encontradas em http://www.inf.pucrs.br/~linatural/index.html.

Partindo de um processo com base linguística e estatística, a principal vantagem dessa ferramenta é que ela fornece, dentre diversas funcionalidades, uma lista dos sintagmas nominais (SNs) que são os mais relevantes de um *corpus* em língua portuguesa, considerando outros *corpora* como elementos de contraste ao *corpus* estudado. Esse processo auxilia a identificação de temas e termos recorrentes e de maior especificidade para o *corpus* em questão. No caso do CLG, especialmente pelos problemas de interpretação relacionados a essa obra, esse trabalho é de grande interesse, no sentido de identificar uma *terminologia saussureana*.

Do ponto de vista prático, para iniciar qualquer trabalho de processamento automático de texto, é preciso preparar os *corpora*. Em primeiro lugar, de modo geral, é preciso obter o texto em formato .pdf, para então converter o arquivo no formato .txt. Após a correção dos erros de grafia que são gerados nessa conversão e da supressão, tanto dos caracteres não processáveis pela máquina, como travessões, quanto das partes irrelevantes para a pesquisa, como a numeração das páginas, pode-se, então, submeter o texto para o processamento automático. Essas decisões já devem ser feitas em conjunto pelo especialista em Computação e em Linguística, sob pena de o recorte não apresentar dados adequados após o processamento.

Para o processamento no ExATOlp, em especial, além da preparação do *corpus* em .txt, é preciso também proceder ao *parsing* do texto, isto é, identificar cada uma das partes do discurso, etiquetando-as, para que o ExATOlp possa reconhecer os SNs e processá-los de modo a apresentar os principais candidatos a SNs relacionados a conceitos importantes no *corpus* a ser analisado. Conforme já salientamos, lidamos com um texto em português que é fruto de tradução do francês, e essa condição tende a repercutir sobre seu tratamento como *corpus* em português.

O *parser* eleito para a etiquetagem prévia de nosso *corpus* em português foi o PALAVRAS, ferramenta desenvolvida por Eckhard Bick, desde 2000, na Universidade de Arhus, Dinamarca. Essa escolha foi guiada pelo fato de que esse é o *parser* atualmente

compatível com o ExATOlp. Dito de modo simples, para que o ExATOlp reconheça os sintagmas mais relevantes no texto com que trabalha, é preciso que a classe e a função de cada uma das palavras que o compõe estejam previamente marcadas no *corpus*.

O processo de anotação linguística do PALAVRAS é aplicado individualmente a cada frase do documento a ele submetido. Cada frase reconhecida é armazenada como uma estrutura em árvore composta por nós terminais, representados graficamente pelas folhas da árvore, onde aparecem as palavras; e por nós não-terminais, onde aparecem as categorias gramaticais. Um exemplo do que faz automaticamente o PALAVRAS, numa versão de uso livre, pode ser visto na Figura 1.

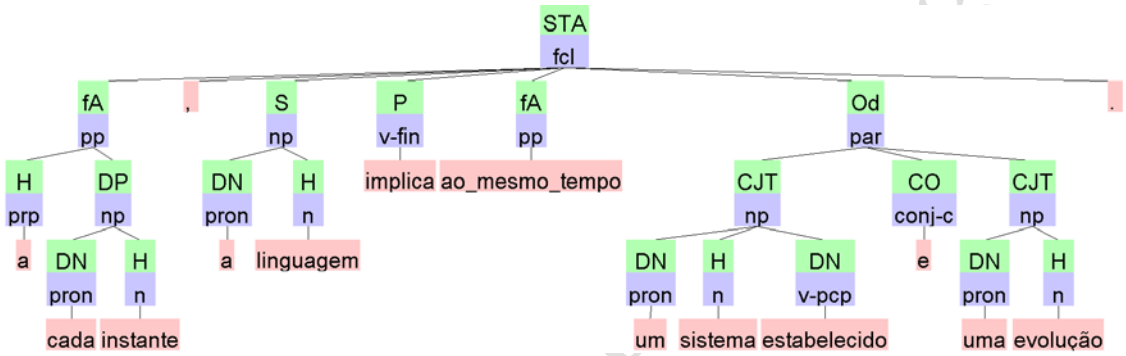

Figura 1. Anotação sintática gerada pelo PALAVRAS para a frase do CLG: A cada instante, a linguagem implica ao mesmo tempo um sistema estabelecido e uma evolução.

Feita a etiquetagem do *corpus* do CLG e do *corpus* de Pneumopatias Ocupacionais pelo PALAVRAS, obtemos arquivos prontos para serem processados pelo ExATOlp. No Diagrama 1 a seguir, em detalhes de cada etapa, o processamento do ExATOlp pode ser mais bem compreendido:

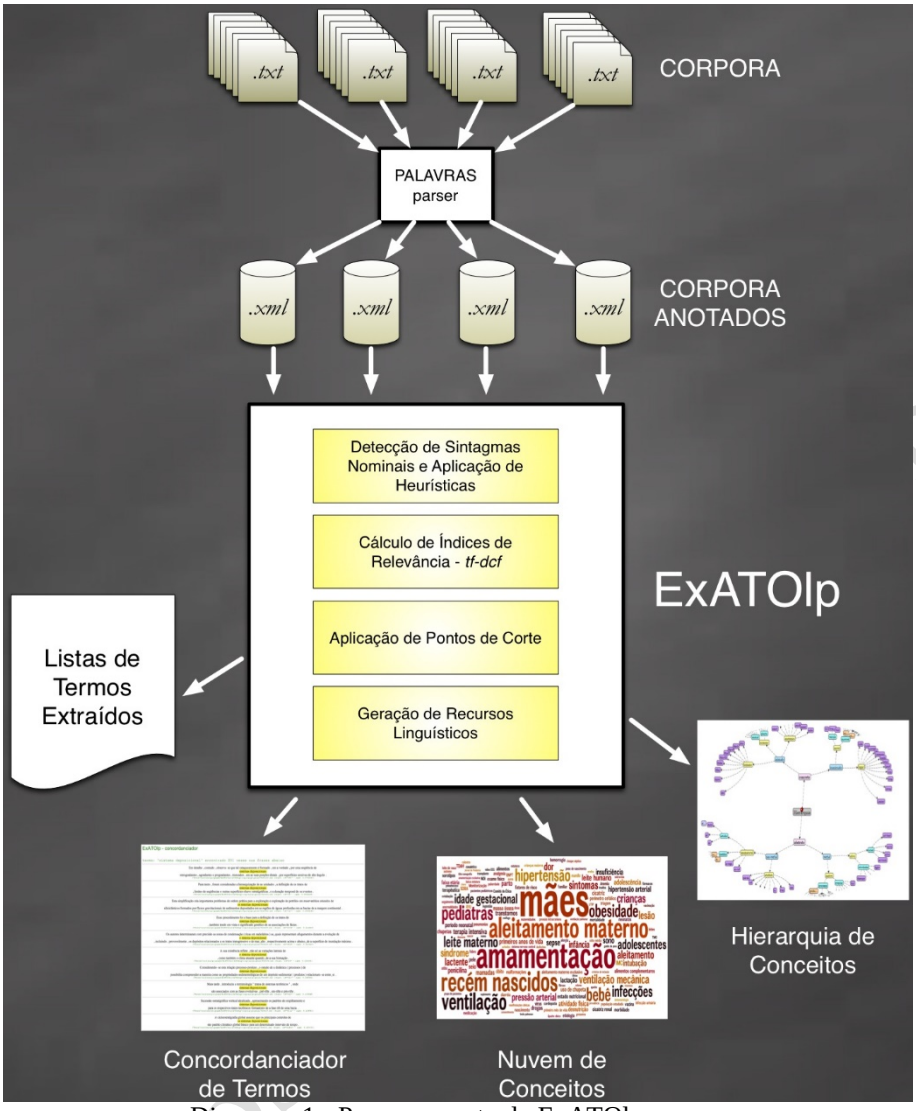

Diagrama 1 - Processamento do ExATOlp.

Cabe salientar que a identificação dos sintagmas nominais mais relevantes feita pelo ExATOlp é um processo com base linguística e estatística. A base linguística manifesta-se por meio de heurísticas que permitem identificar com maior qualidade os SNs do *corpus*, enquanto a base estatística se manifesta por meio do cálculo de um índice de relevância que leva em conta a frequência e especificidade de cada SN extraído, sempre em relação aos *corpora* ou ao *corpus* de contraste. O contraste é feito, comparando-se os índices de frequência dos SN do *corpus* de estudo com os que são encontrados em outros *corpora*, entendido, aqui, que todos os SN, tanto do *corpus* de estudo como os dos *corpora* de contraste, foram previamente selecionados pelo sistema do ExATOlp entre os que têm maior chance de desempenhar um papel de termo, pela sua posição sintática (de sujeito ou objeto). Assim, *grosso modo*, após essa seleção, se um desses SN é bastante frequente no *corpus* de estudo, mas também nos *corpora* de contraste, isso significa que

não é um bom candidato a termo, pois não é específico da área de domínio do *corpus* em estudo. Por outro lado, se ele aparece pouco ou nem aparece nos *corpora* de contraste, mas aparece no *corpus* de estudo, é um candidato a termo. Informações mais detalhadas do processo com base linguística pode ser encontrado no trabalho de Lopes (2012), e do processo com base estatística, no trabalho de Lopes; Fernandes; Vieira (2012).

A seguir, nas Figuras 2 e 3, o resultado do processamento do CLG, a partir do contraste com *corpora* das áreas de Pediatria, Computação e Geologia, é representado graficamente através de uma das possibilidades que o ExATOlp oferece para os resultados. Trata-se de uma árvore hiperbólica, trazendo a partir do centro os termos mais relevantes do *corpus* a ele submetido. As árvores hiperbólicas podem ser movimentadas e aumentadas, o que facilita a visualização dos SNs e de suas relações com outros SNs. Porém, aqui nas figuras é apenas uma ilustração.

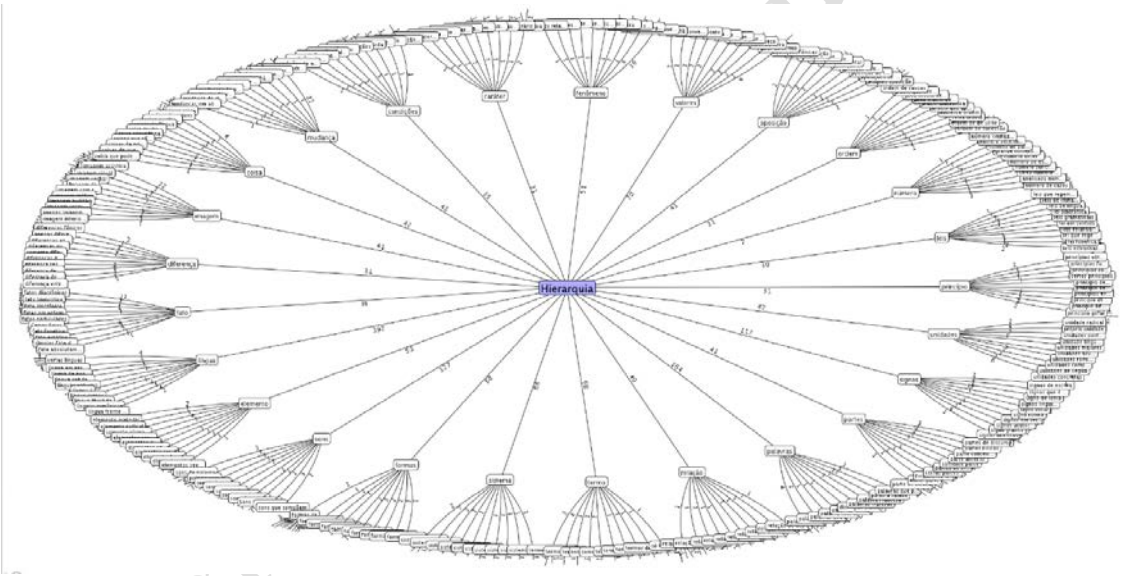

Figura 2. Visualização da hierarquia de conceitos do CLG extraída pelo ExATOlp.

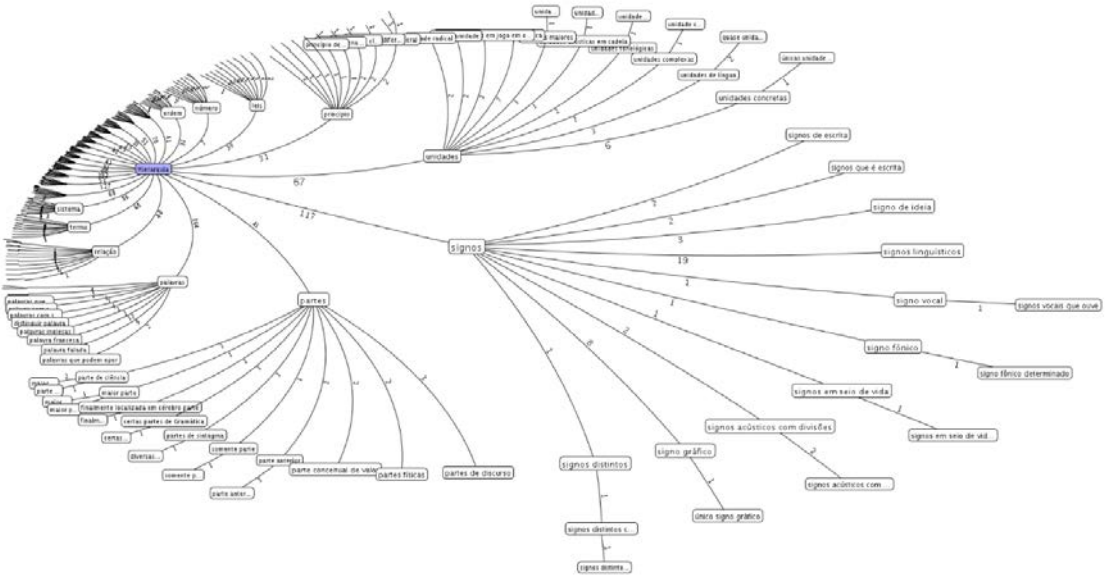

Figura 3. Visualização da hierarquia de conceitos do CLG extraída pelo ExATOlp, com ênfase nas relações de um dos termos (o termo “signos”).

Nas árvores hiperbólicas, os termos aparecem em etiquetas, ligados aos outros termos aos quais se relacionam semanticamente, o que a autora chama de *hierarquia de conceitos*. A árvore hiperbólica gerada pelo ExATOlp permite uma visualização interativa, ou seja, é possível clicar nas etiquetas, modificando o formato do gráfico, para visualizar em detalhe as relações entre cada um dos termos, conforme demonstramos na Figura 3. Além disso as arestas que conectam um termo do nível mais alto da hierarquia com seu sucessor imediato indicam o número de ocorrências desse termo sucessor no *corpus*. Por exemplo, o termo “signos” está ligado ao termo “signos linguísticos” por uma aresta que indica 19 ocorrências do termo “signos linguísticos”. Igualmente a aresta que liga o centro da árvore (etiqueta “Hierarquia”) ao termo “signos” indica que esse termo ocorreu 117 vezes no *corpus*. Cabe salientar que o número de ocorrências expresso na árvore hiperbólica serve como indicativo da importância de cada termo para o domínio.

Outra possibilidade de visualização das informações extraídas pelo ExATOlp é a *nuvem de conceitos*, como mostramos na Figura 4, que indica os termos mais relevantes do *corpus* CLG. Nesse tipo de representação, o tamanho da letra com que os SNs são escritos varia de acordo com sua relevância no *corpus*, da letra maior, para o mais relevante, à menor, para o menos relevante. Conforme citado anteriormente, a relevância leva em consideração a frequência e especificidade de cada termo em comparação com os *corpora* de contraste. Dessa forma, a Figura 4 indica que os termos “Língüística” e “signos” são claramente mais relevantes que todos os demais.

Figura 4. Visualização da nuvem de conceitos do CLG extraída pelo ExATOlp, com os SNs mais relevantes em letras maiores.

De acordo com os objetivos de análise, o ExATOlp possibilita a visualização em separado dos termos, segundo seu número de palavras. As Figuras 5, 6 e 7 apresentam respectivamente os unigramas, bigramas e trigramas mais relevantes. Esse tipo de análise em separado possibilita estudos mais específicos em relação a termos compostos que podem ser observados com mais clareza quando separados dos unigramas.

Figura 5. Visualização da nuvem de unigramas mais relevantes do CLG extraída pelo ExATOlp.

Figura 6. Visualização da nuvem de bigramas mais relevantes do CLG extraída pelo ExATOlp.

Figura 7. Visualização da nuvem de trigramas mais relevantes do CLG extraída pelo ExATOlp.

Observamos também que, na listagem de bigramas, aparecem mais termos que podem ser reconhecidos como específicos da Linguística Saussureana, como "imagem acústica", "signos linguísticos", "Linguística diacrônica" e "relação sintagmática". E, com raras exceções, como "igual modo", todos os bigramas mais frequentes são termos específicos do domínio da Linguística. Nos trigramas, são colocados em relevo outros tantos termos importantes e específicos da teoria de Saussure, como "estado de língua", "arbitrariedade do signo" e "sistemas de valores". Aparentemente, bi e trigramas são menos propensos à ambiguidade do que unigramas. A combinação de termos simples em compostos concede especificidade ao conceito, caracterizando o SN como termo. Além disso, há uma tendência, e essa é uma hipótese a ser testada em trabalhos futuros, de que termos compostos sejam mais frequentes em textos especializados.

Apesar de ainda inicial, esse estudo revela que a aplicação de métodos automáticos para a extração de informação pode ser muito útil para o trabalho do terminólogo, identificando de maneira rápida itens lexicais que podem ser considerados efetivamente como termos. Essa mesma tarefa, considerando-se o trabalho humano, demandaria muito mais tempo, especialmente numa obra complexa como o CLG.

Uma outra observação importante é a de que a maior parte dos termos considerados como relevantes pelo ExATOlp dizem respeito à Fonologia. Esse resultado, que aponta para uma saliência dos estudos em Fonologia a partir de Saussure, não traz exatamente uma novidade, mas pode ser considerado surpreendente, já que outros assuntos, como a arbitrariedade do signo e as dicotomias (entre língua e fala, associações sintagmáticas e paradigmáticas, sincronia e diacronia, por exemplo) ocupam, em geral, a posição dos "assuntos saussureanos por excelência". A relevância da contribuição de Saussure, no CLG, para a Fonologia fica, então, como sugestão para pesquisas futuras. Tal observação confirma, em parte, a hipótese inicial de nosso trabalho, de que uma visão

panorâmica da obra, via ExATOlp, poderia proporcionar uma visão diferenciada da obra, que normalmente é estudada por partes, de modo fragmentado.

Além da análise dos resultados obtidos a partir dessas duas diferentes metodologias, para um trabalho futuro, será elaborada a lista de referência dos principais termos relacionados a conceitos importantes em Saussure, a partir da opinião de especialistas, o que é interessante do ponto de vista do aprofundamento da reflexão teórica e é necessário para o teste de precisão de ferramentas, em especial o ExATOlp e diferentes metodologias.

Outra pesquisa que se apresenta como sugestão, a partir dos resultados da extração automática de termos, é sobre as recategorizações, ou seja, com que termos e tipos de construções os conteúdos foram referidos e retomados e quais as consequências dessas escolhas. Ainda que se trate de um trabalho com resultados mais voltados para os estudos linguísticos do texto, ele pode ser útil no sentido de aperfeiçoar sistemas de extração automática e a análise de seus resultados.

4. Considerações finais

Neste artigo, mostramos que, se de um lado, a Computação e o PLN trabalham com sistemas de organização da informação e dados estatísticos, visando criar ferramentas computacionais com melhor desempenho no tratamento da linguagem, a Linguística, por outro lado, pode fornecer a descrição de regras ou de alguns padrões para que se possa lidar com dados linguísticos dentro de um *corpus* previamente selecionado e preparado. Note-se que a própria seleção e preparação eficiente desse material, assim como todo o processo de tratamento automático de fenômenos linguísticos, depende do conhecimento de ambas as partes.

Apresentamos também um exemplo prático e bem-sucedido de cooperação entre pesquisas de PLN e Estudos Linguísticos. Essa é uma frente de investimentos promissora, que tende a evidenciar e a divulgar o que faz um linguista e que importância tem o seu trabalho em diferentes contextos de aplicação, especialmente no que diz respeito ao tratamento da língua portuguesa em ambientes digitais.

Cabe salientar que a geração de todos os resultados apresentados nas figuras desse artigo foram gerados de maneira completamente automática pelo ExATOlp sem necessidade de intervenção humana. No entanto, esses resultados automáticos carecem de uma profunda análise humana que só linguistas estão capacitados a executar. Esse exemplo ilustra, portanto, o potencial colaborativo do PLN e Estudos Linguísticos.

Por fim, consideramos importante divulgar uma pesquisa como a nossa, que toma o CLG como *corpus.* O relato visa incentivar outros colegas linguistas, especialmente os que já tenham tido contato com algum produto Linguística de *Corpus*, a avançar na direção do diálogo com os colegas do PLN, especialmente no Brasil, onde nos deparamos com uma carência de estudos que associam linguagem e tecnologia. Hoje, não se trata mais de uma opção, mas sim, de uma necessidade.

Referências bibliográficas

BERBER SARDINHA, T. **Linguística de Corpus**. Barueri, SP: Manole, 2004.

BRANCO, A.; MENDES, A.; PEREIRA, S.; HENRIQUES, P.; PELLEGRINI, T.; MEINEDO, H.; TRANCOSO, I.; QUARESMA, P.; STRUBE DE LIMA, V. L.; BACELAR, F. **The Portuguese Language in the Digital Age / A Língua Portuguesa na Era Digital**. 1. ed. Berlin: Springer, 2012. v. 1.

BOUQUET, S. **Introdução à leitura de Saussure**. São Paulo: Cultrix, 2000.

CIULLA, A. ; FINATTO, M. J. B. O signo linguístico em Saussure: algumas questões sobre a tradução para o português brasileiro. **Traduzires**, v. 2, p. 55-64, 2013.

CULLER, J. **As idéias de Saussure**. São Paulo: Cultrix, 1979.

DIAS-DA-SILVA, B. C. **A fase tecnológica dos estudos da linguagem**: o processamento automático das línguas naturais. Tese (Doutorado em Linguística e Língua Portuguesa) – Faculdade de Ciências e Letras, UNESP, Araquara, 1996.

_____. Concepções e finalidades da análise gramatical e o processamento automático das línguas naturais. In: SEMINÁRIO DO GRUPO DE ESTUDOS LINGUÍSTICOS DO ESTADO DE SÃO PAULO (GEL), 1998, São José do Rio Preto. **Programação e Resumos...** São José do Rio Preto: Setor de Publicações/IBILCE-UNESP, 1998, v. XLV, p.185-185.

_____. A construção da base da Wordnet.Br: conquistas e desafios. In: III WORKSHOP DE TECNOLOGIA DA INFORMAÇÃO E DA LINGUAGEM HUMANA (UNISINOS), 2005, São Leopoldo. **Anais...** São Leopoldo: Unisinos, 2005, p.2238-2247.

JÄGUER, L. La pensée épistémologique de F. de Saussure. In: BOUQUET, S. (Ed.). **L'Herne**: Saussure. Paris: L'Herne, 2003.

LOPES, L. **Extração automática de conceitos a partir de textos em língua portuguesa**. 2012. 113f. Tese (Doutorado em Ciência da Computação) - Faculdade de Informática, PUCRS, Porto Alegre, RS, 2012.

LOPES, L., FERNANDES, P., VIEIRA, R. Domain term relevance through tf-dcf. In: ICAI - International Conference in Artificial Inteligence, 2012, Las Vegas, EUA. **Proceedings of ICAI'12**. Las Vegas, EUA: Worldcomp, 2012. p. 1-7.

MARTINS, R. O pecado original da linguística computacional. **Revista Alfa**, São Paulo, v.55, n.1. 2011.

NORMAND, C. **Saussure**. São Paulo: Estação da Liberdade, 2009.

RAJOGOPALAN, K. A linguística de *corpus* no tempo e no espaço: visão reflexiva. In: GERBER, R.M.; VASILÉVSKI, V. (Orgs.). **Um percurso para pesquisas com base em corpus**. Florianópolis, SC: Editora da UFSC, 2007, p.23-44.

ROSA, J.L.G. **Fundamentos da Inteligência Artificial**. Rio de Janeiro: LTC, 2011.

TRABANT, J. Faut-il défendre Saussure contre ses amateurs? Notes item sur l'étymologie saussurienne. In: CHISS, Jean-Louis; DESSONS, Gérard. **Langages**. Paris: Larousse, n.159, septembre, 2005.

Artigo recebido em: 09.01.2015
Artigo aprovado em: 05.06.2015