DOI: 10.14393/DL52-v16n4a2022-1

Tratamento computacional do português brasileiro

Computational treatment of Brazilian Portuguese

Heliana MELLO[*]
Fernanda FARINELLI[**]

1 O tratamento computacional das línguas naturais

O tratamento computacional de dados linguísticos tem estado na agenda de linguistas e cientistas da computação há no mínimo cinco décadas; entretanto, apenas nas últimas duas décadas tal movimento ganhou impulso no cenário brasileiro. Este movimento conta com a adesão de pesquisadores de diversas áreas do conhecimento, que progressivamente, através das novas tecnologias e formações acadêmicas mais sintonizadas com as necessidades do tratamento de línguas naturais via procedimentos computacionais, vão ganhando visibilidade.

É relevante que destaquemos aqui o quão importante a formação dos jovens graduandos, sobretudo na área de estudos linguísticos, esteja alinhada às pautas de pesquisa e inovações metodológicas que a área de tratamento computacional de línguas naturais exige. Por isso, somos fortes defensoras do ensino de programação e estatística na formação linguística e da promoção de interação com os conhecimentos oriundos das áreas informáticas e da computação.

A linguística de *corpus* de chegou ao Brasil há cerca de duas décadas, à época, com uma predominância de atuações voltadas para as subáreas do ensino de línguas estrangeiras, sobretudo língua inglesa, e estudos da tradução. Os estudiosos de

[*] Doutora em Linguística (CUNY, EUA), Professora Titular da Faculdade de Letras, UFMG. ORCID: https://orcid.org/0000-0003-0267-9005. hmello@ufmg.br
[**] Doutora em Ciência da Informação (ECI-UFMG), Professora Adjunta da Faculdade de Ciência da Informação, UnB. ORCID: https://orcid.org/0000-0003-2338-8872. fernanda.farinelli@unb.br

fraseologia e lexicografia também se envolveram com os estudos de *corpora* e formação de bancos de dados linguísticos.

No Brasil temos pesquisadores que poderiam ser chamadas de precursores da linguística de *corpus* e do tratamento computacional do português brasileiro (PB). Na frente da linguística, pesquisadores como Tony Berber Sardinha, Violeta Quental, Bento Dias da Silva, Leonel Figueiredo, dentre muitos outros; bem como as equipes do NILC USP-São Carlos, da UFRS e da PUC-RS, dentre outras, na frente da Ciência da Computação e da Informática, nos vêm imediatamente à mente quando tentamos traçar o percurso do tratamento computacional do PB.

A tradição criada pelos nossos pioneiros é generosa, com o compartilhamento de muitos *corpora* e ferramentas computacionais gratuitos e passíveis de serem utilizados pela comunidade livremente.

Para além dos pioneiros brasileiros, devemos mencionar o portal LINGUATECA, liderado por Diana Santos, que baseado em Portugal, tem disponibilizado gratuitamente, ao longo de décadas, tanto *corpora* quanto ferramentas computacionais para o estudo de diversas variedades do português, inclusive o PB. Agregando usabilidade e camadas de anotação a diversos *corpora* disponíveis na LINGUATECA, está o *parser* PALAVRAS, desenvolvido por Eckhard Bick (BICK, 2000).

A nossa intenção ao propor a temática deste volume à Revista Domínios de Lingu@gem foi a de oferecer aos leitores um mapa, mesmo que incompleto, da riqueza da produção da área e a necessidade cada vez mais urgente de formarmos jovens aptos a lidar com as grandes perguntas metodológicas, que se colocam para o linguista e o cientista da computação, e outras formações profissionais, que desejam acompanhar o inexorável fluxo do desenvolvimento tecnológico e seu impacto na pesquisa linguística. Todos nós, que trabalhamos na área, compartilhamos bases empíricas e

metodológicas complexas e múltiplas, que exigem o trabalho em colaboração e a troca rápida e eficiente de saberes.

O tratamento computacional das línguas naturais pode, grosso modo, ser subdivido em três grandes campos que não guardam, absolutamente, fronteiras rígidas entre si:[1] compilação e tratamento de *corpora*, estudos pautados pela linguística computacional (metodologias para o tratamento computacional de línguas naturais) e processamento de linguagem/língua natural. Esses três campos contam com profissionais de diversas áreas do saber, mas a colaboração entre linguistas e cientistas da computação é a interface de competências predominante. Para os dois primeiros campos, há uma primazia se não de liderança de linguistas, pelo menos de perguntas de pesquisa de ordem linguística, já que o foco principal, mesmo que intermediado por código e ferramentas computacionais, é linguístico; enquanto no terceiro, a liderança geralmente é de cientistas da computação, interessados cada vez mais em inteligência artificial e seus métodos, que exigem *big data* para extrair informações de línguas naturais. Este número temático seguirá essa divisão tripartite, admitidamente difusa, no grupamento de seus doze artigos.

A base comum às três subáreas mencionadas acima é, obviamente, a utilização de métodos computacionais para o tratamento das línguas naturais. Tal tratamento pode lidar com dados linguísticos não-estruturados (textos, como os conhecemos), semiestruturados (representação heterogênea, textos formatados via XML ou JSON, por exemplo) ou estruturados (em formato de tabela, como num banco de dados relacional). Igualmente, a intervenção humana no tratamento de tais dados varia entre dois polos: intervenção humana ativa (como no processo de anotação manual, mesmo que com utilização de uma interface computacional) e processo totalmente

[1] Essa subdivisão não é acatada por todos que atuam na área. Há visões distintas sobre o assunto e a nossa proposta parte de uma preocupação pedagógica sobre o tratamento computacional de línguas naturais, sem uma adesão rígida a qualquer corrente teórica da área. Ver Othero e Menuzzi (2005, p. 17) para uma visão que divide a linguística computacional entre linguística de corpus e PLN.

automatizado e conduzido pelas máquinas (como em aprendizado profundo ou redes neurais profundas).

O primeiro *corpus* linguístico computacional foi o Brown Corpus of Standard American English[2], compilado em 1961, a partir de uma variedade de gêneros textuais, contendo 1 milhão de palavras e, posteriormente publicado (1964) e apresentado, com aplicações, por Francis e Kučera (1967). Já este trabalho precursor sinalizava um caminho de muito interesse para a compilação de dados linguísticos habilitados ao tratamento computacional e à sua usabilidade para diversos tipos de análises linguísticas.

Outro *corpus* de importância histórica no cenário internacional é o London-Lund Corpus of Spoken British English[3], contendo cerca de 500 mil palavras (SVARTVIK 1990). Pela primeira vez, dados de fala foram coletados e compilados em um formato de *corpus*. Este é um *corpus* incluído no que se chama usualmente de "primeira onda da linguística de *corpus*", em que *corpora* orais disponibilizavam as transcrições da fala, ainda sem a possibilidade de se conectar o sinal sonoro à sua correspondente transcrição gráfica.

Outro passo significativo na área foi a ampliação nos tamanhos dos *corpora* compilados, passando-se à escala dos milhões e, posteriormente, bilhões de palavras. Num momento histórico em que *big data* tornou-se uma colação frequente em diversas áreas do saber, também nos estudos de *corpora* a escalada dos dados é algo que vem ocorrendo e que se faz necessário em diversos tipos de aplicações linguísticas, como no exemplo clássico da composição de dicionários, além de ser crucial para tarefas computacionais variadas, realizadas através de aprendizado de máquina na área de processamento de língua natural (PLN).

[2] http://icame.uib.no/brown/bcm.html
[3] http://korpus.uib.no/icame/manuals/LONDLUND/INDEX.HTM

A linguística de *corpus* ganhou, desde os seus anos iniciais, uma enorme variedade de métodos para a compilação de dados, os quais, se inicialmente eram representativos da língua escrita, ampliaram-se para incluir a língua falada e a gestualidade, havendo hoje os *corpora* multimodais, em que gestos, fala e transcrição são alinhados sincronamente, permitindo ao pesquisador uma experiência realisticamente empírica na análise de dados.

Para além de diamesias variadas, os *corpora* disponíveis atualmente cobrem um impressionante espectro de épocas históricas, áreas de especialização, línguas, propósitos de pesquisa, formatos, tipologias de anotação e escalabilidade.

Esta mesma tendência pode ser observada nos *corpora* dedicados ao estudo do PB, conforme os exemplos a seguir. O Corpus Brasileiro[4], traz textos de diversas tipologias, inclusive textos transcritos de fala, perfazendo 1 milhão de palavras, podendo ser acessado gratuitamente através do portal LINGUATECA[5] ou através do seu site próprio (ver nota 4). O portal de ferramentas e recursos do NILC[6] também oferece uma grande variedade de *corpora* escritos, além de ferramentas computacionais e bases de dados para PLN. O Corpus NILC, com mais de 40 milhões de palavras, é disponibilizado através do portal do NILC e também da LINGUATECA[7]. O PB conta hoje com *corpora* orais com alinhamento síncrono sinal sonoro-transcrição, como o C-ORAL-BRASIL[8] e o NURC Digital[9], além de *corpora* especializados voltados para a tradução, a fraseologia, a lexicografia, o ensino de línguas, dentre outras aplicações, como pode ser apreciado no Portal do Projeto COMET[10].

[4] http://corpusbrasileiro.pucsp.br/cb/Inicial.html
[5] https://www.linguateca.pt/acesso/corpus.php?corpus=CBRAS
[6] http://www.nilc.icmc.usp.br/nilc/index.php/tools-and-resources
[7] https://www.linguateca.pt/acesso/corpus.php?corpus=SAOCARLOS
[8] https://www.c-oral-brasil.org/
[9] https://fale.ufal.br/projeto/nurcdigital/
[10] https://comet.fflch.usp.br/projeto

Passando agora, à linguística computacional, tomamos aqui a visão de Freitas (no prelo, p. 4)[11], que entende que o termo linguística computacional pode ser utilizado quando a dimensão linguística de uma pesquisa está em evidência, mesmo que o trabalho computacional que a subjaz esteja muito alinhado com PLN. Assim, a linguística computacional desenvolve métodos computacionais para responder a perguntas científicas da linguística. Caberiam aqui trabalhos que tentam formalizar e investigar, através de metodologias computacionais, questões relacionadas ao conhecimento linguístico, aquisição e uso linguístico, distribuição do léxico em estruturas linguísticas, gramáticas para dados de língua natural, dentre outros. Destacam-se neste escopo, adicionalmente, ferramentas computacionais desenvolvidas especificamente para atender às necessidades analíticas oriundas da linguística. Num certo sentido, como compartilhado por Jason Eisner[12], bons modelos em linguística computacional, que mais e mais deem conta da competência linguística humana, acarretarão melhores modelos em PLN, se estes se basearem nos frutos gerados pela pesquisa que leva a sério as perguntas fundamentais da investigação linguística.

Os avanços em linguística computacional internacionalmente, assim como aqueles em PLN, são tantos e tão rápidos, que não caberia discuti-los aqui. Dentro da visão que adotamos para a linguística computacional, é válido, entretanto, mencionar ferramentas ou conjuntos de ferramentas computacionais que têm grande empregabilidade nos estudos linguísticos. Dentre eles, os anotadores PoS e os *parsers* sintáticos são provavelmente as ferramentas mais bem conhecidas e absolutamente necessárias para o trabalho com *corpora*. Adicionalmente os lematizadores e *stemmers* são ferramentas básicas à investigação linguística computacional. Mas, na verdade,

[11] A autora considera linguística computacional aproximadamente como PLN, que por sua vez é parte de inteligência artificial, logo, um campo dentro da ciência da computação (*op. cit*., p. 3).
[12] https://www.cs.jhu.edu/~jason/

qualquer tipo de anotação automática com alto nível de acurácia, seja ela morfossintática, semântica ou morfológica é extremamente útil no tratamento computacional de dados linguísticos.

Para além dos diversos tipos de anotadores, as ferramentas atualmente desenvolvidas para a identificação de elementos de impacto linguístico que transcendam o texto escrito, como o software de análise acústica PRAAT[13], também ocupam um espaço crucial nas investigações linguísticas.

Conjuntos de ferramentas gratuitas, desenvolvidas por linguistas computacionais, como as oferecidas por Lawrence Anthony[14], são de grande utilidade e acessibilidade, auxiliando tanto linguistas, quanto professores e estudantes de línguas.

Para o tratamento computacional do PB há um longo histórico de ferramentas computacionais oferecidas pela equipe do NILC[15], para além das dissertações e teses desenvolvidas nesse núcleo de pesquisa. Também de destaque são iniciativas de desenvolvimento de ferramentas, de pesquisadores que não necessariamente atuam em equipes institucionais de grande porte, como as de Leonel Figueiredo de Alencar Araripe[16], da UFC, desenvolvedor do *parser* Aelius[17]. Notória também é a contribuição de Plínio de Almeida Barbosa[18] para a pesquisa das ciências da fala, com o desenvolvimento de ferramentas computacionais junto ao seu grupo de colaboradores, como o recente Alinha-PB[19].

Movendo-nos, agora, para a contribuição de PLN, entramos em um universo em que a escalabilidade dos dados e os modelos de aprendizado de máquina, cada vez

[13] https://www.fon.hum.uva.nl/praat/
[14] https://www.laurenceanthony.net/software.html
[15] http://www.nilc.icmc.usp.br/nilc/index.php/tools-and-resources
[16] https://scholar.google.com.br/citations?user=H7HoA0AAAAAJ&hl=pt-BR
[17] https://github.com/LR-POR/aelius
[18] https://www.researchgate.net/profile/Plinio-Barbosa-2
[19] https://conversoralinhador.herokuapp.com/about

mais sofisticados e com resultados impressionantes, se tornam a prática da área. Para além da contribuição de linguistas e cientistas da computação, essa área conta com matemáticos, estatísticos, físicos, engenheiros – sendo, assim, uma área verdadeiramente multidisciplinar.

É historicamente difícil dizer quando foi o início dessa área de pesquisa. Há opiniões divergentes (cf. CHURCH; LIBERMANN, 2021). Mas, uma referência seminal é Turing (1950), quando o famoso Teste de Turing é proposto. Neste teste, um avaliador humano deve decidir se respostas a perguntas em língua natural foram produzidas por um humano ou por uma máquina. No momento presente, 2022, ainda há diferenças significativas na capabilidade de uma máquina emular seja a produção ou a compreensão de uma língua natural. Entretanto é inegável o imenso avanço da área e as inumeráveis aplicações de PLN já alcançados. Os progressos da área se devem a uma conjunção de fatores: a disponibilidade de volumes cada vez maiores de dados linguísticos em formato digital, avanços na pesquisa linguística, crescimento exponencial do poder de computação, e o desenvolvimento de novos métodos, principalmente em aprendizado de máquina.

Tanto no contexto acadêmico quanto no industrial, um grande repertório de tarefas são desenvolvidas, como a classificação de textos, extração de informação, análise de sentimento, tradução automática, síntese de fala, reconhecimento de fala, geração de textos, sumarização, *chatbots* e agentes virtuais, dentre outras em um rol imenso de produtos.

O investimento que a indústria dedica à área pode ser testemunhado facilmente através do acesso a qualquer dos serviços oferecidos pelas empresas tecnológicas internacionais via *world wide web*. Os nossos computadores pessoais tornaram-se um observatório privilegiado para a chamada *BigTech*, que espiona com muito sucesso toda a produção linguística veiculada pelas nossas interações com a máquina, extraindo as informações que lhe são de interesse comercial.

As aplicações de PLN estão também presentes em contextos que afetam a vida da humanidade em todo o planeta, como na medicina, ou na gestão do trânsito nas grandes metrópoles, ou até mesmo nas aplicações forenses, cruciais tantas vezes na resolução de atividades criminosas.

Em um artigo de opinião recente, Kavinski (2022) afirma que podemos estar próximos do momento em as línguas naturais deixarão de ser exclusivamente humanas e passarão a fazer parte da nossa interação com as máquinas e os objetos. Parece-nos bastante ousada essa afirmação, mas é fato que não sabemos qual será o limite para PLN e questões éticas têm sido progressivamente mais discutidas pelos praticantes da área (cf. os recursos disponibilizados em https://aclweb.org/aclwiki/Ethics_in_NLP).

No contexto brasileiro, acompanhando a tendência internacional, sabemos que há interesse e investimento em PLN pelas grandes empresas que atuam em diversos setores de serviços. Isso, interessantemente, tem criado novas possibilidades de trabalho para jovens linguistas já formados dentro de programas de estudo que consideram a programação como algo necessário para a atuação profissional.

Na academia, praticamente em todas as universidades em que há núcleos de pesquisa em ciência da computação, informática, matemática e estatística aplicada, há também pesquisadores envolvidos com projetos em PLN, muitas vezes voltados para aplicações industriais, médicas ou comerciais. A produção científica na área cresce logaritmicamente, passando por um momento de muita efervescência e projetos de muito interesse social. A título de ilustração, sugerimos a verificação de artigos listados pela OMS relacionados ao coronavírus, que envolvem PLN como base metodológica (cf. https://search.bvsalud.org/global-literature-on-novel-coronavirus-2019-ncov/?lang=pt&q=mh:%22Processamento%20de%20Linguagem%20Natural%22).

No contexto brasileiro, mencionaremos apenas um grande projeto que nos parece particularmente relevante quando se discute PLN para o português brasileiro:

o NLP2[20], liderado por Marcelo Finger e Thiago Pardo. Esse projeto guarda-chuva faz paz parte do Center for Artificial Intelligence e agrega pesquisadores de diversas instituições, com o propósito de desenvolver recursos que avancem PLN em português. Dentre suas frentes de trabalho estão o processamento do português via análise sintática e *parsing*, tarefas de anotação para reconhecimento e síntese da fala e a criação de um *corpus* geral do português brasileiro contemporâneo.

Temos hoje no Brasil equipes que trabalham desenvolvendo recursos e ferramentas linguístico-computacionais que acompanham o estado da arte no cenário internacional – uma tendência que, esperamos, continue a se desenvolver e a apoiar cada vez mais a nossa compreensão da língua portuguesa e das aplicações de impacto social que a partir dela podem ser criadas.

Também crescente é o número de associações profissionais e eventos acadêmicos (cf. STIL, JDP) que têm dado espaço a trabalhos voltados para o processamento computacional do português brasileiro.

Por fim, gostaríamos de mencionar o grupo Brasileiras em PLN[21], fundado em 2020 por Helena Caseli e Brielen Madureira. O grupo reúne mais de 100 mulheres brasileiras que atuam em PLN sob vários formatos de inserção, e tem uma agenda que destaca a atuação das mulheres brasileiras neste campo científico.

Passaremos, a seguir, à apresentação dos doze artigos que compõem este volume temático e que foram agregados de acordo com as três vertentes temáticas discutidas nesta seção 1.

2 Estudos sobre Tratamento Computacional do Português Brasileiro

Os artigos apresentados a seguir, de uma maneira ou de outra abordam questões relativas a *corpora* do português brasileiro, seja ao tratar aspectos da

[20] https://c4ai.inova.usp.br/pt/pesquisas/#NLP2_port
[21] https://sites.google.com/view/brasileiras-pln/in%C3%ADcio

composição de *corpora*, sua anotação e sua exploração para a abordagem de problemas linguísticos específicos, seja no tratamento via ferramentas computacionais de dados de *corpora* ou na exploração de algoritmos e técnicas de aprendizado de máquina. Os *corpora*, como discutido na ampla literatura da área, são a base e o ponto de partida para o tratamento computacional das línguas naturais. Alguns dos artigos trazem revisões bibliográficas sobre o desenvolvimento da linguística de *corpus*, da linguística computacional e do processamento de línguas naturais. Há visões distintas sobre as áreas envolvidas, bem como adoção terminológica variada.

Conforme já mencionado, com o propósito de melhor orientar o leitor, os doze artigos que compõem este volume foram agrupados em três eixos temáticos principais: (a) compilação, tratamento e exploração de *corpora*, (b) estudos pautados pela linguística computacional e (c) processamento de língua/linguagem natural (PLN). Os cinco primeiros trabalhos (Santos, Kauffmann, Gomide *et al*., Cruz, Ferrari e Cunha) representam o eixo sobre aplicações de linguística de *corpus*. O segundo eixo temático, metodologias da linguística computacional para a análise linguística, aglutina mais cinco artigos (Rodrigues *et al*., Raso *et al*., Santos Junior e Isquerdo, Alencar e Rademaker, Bick). E, finalmente, o terceiro eixo temático, sobre processamento de língua natural (PLN) reúne os dois trabalhos finais deste volume (Duran *et al*., Souza e Di Felippo).

O volume é aberto pelo estudo de Diana Santos (Universidade de Oslo e Linguateca) intitulado **"A Gramateca e a Literateca como macroscópios linguísticos"**. Este estudo apresenta uma breve contextualização sobre os ambientes da Gramateca e da Literateca, que integram a Linguateca, trazendo exemplos de perguntas de pesquisa acerca do português que estes ambientes são capazes de responder. A autora apresenta uma discussão sobre as diferentes potencialidades e funcionalidades destes ambientes para apoiar a pesquisa em língua portuguesa, principalmente o português brasileiro, demonstrando o papel destes ambientes como um observador macroscópio da língua

do ponto de vista semântico e morfossintático. Adicionalmente reflete sobre estes ambiente no ensino do português, na leitura distante de textos literários e na extração de informação em língua portuguesa.

O segundo artigo intitulado **"Cognição e variação linguística de gêneros/registros jornalísticos: um estudo baseado em corpus"** escrito por Carlos Henrique Kauffmann (PUC-SP) investiga o processo de identificação e reconhecimento de gêneros/registros jornalísticos na diferenciação entre textos. Para isso, o autor descreve sua metodologia que combina os recursos metodológicos da linguística de *corpus* a uma pesquisa qualitativa, a fim de medir o grau de concordância de um grupo de especialistas quanto aos gêneros/registros existentes nos textos jornalísticos. O *corpus* avaliado foi formado por textos de duas edições do jornal "Folha de São Paulo" impresso, totalizando 1.431 textos. Os resultados obtidos foram tabulados pelo grau de concordância entre os classificadores, levando em conta as dimensões do texto jornalístico identificadas pelo autor em um estudo de 2005, demonstrado uma tendência das classificações para as categorias reportagem e notícia.

Na sequência, as autoras Andressa Rodrigues Gomide (Universidade de Coimbra), Taíse Simioni (UNIPAMPA) e Aden Rodrigues Pereira (UNIPAMPA) exploram "**O fenômeno do desfocamento do agente uma discussão sobre a importância dos recursos computacionais para os estudos linguísticos"**. O artigo, trata das possibilidades de uso de ferramentas da linguística de *corpus* (LC) e de um *corpus* de escrita acadêmica em português (CoPEP) para explorar o fenômeno do desfocamento do agente em artigos acadêmicos publicados no Brasil e em Portugal. As autoras discutem o papel da LC para análises linguísticas sintetizando alguns dos seus aspectos teóricos e pressupostos mais importantes para medir a precisão da ocorrência do fenômeno linguístico, foco de sua pesquisa. Para sustentar a proposta de diferenciação entre as construções frasais e seu lugar no contínuo de desfocamento do agente, as autoras apresentam resultados de dois estudos empíricos sobre o tema

para o português brasileiro (PB) e suas implicações. Por fim, o tema desfocamento do agente em um *corpus* de escrita acadêmica é abordado através da descrição do processo de preparação do *corpus* (tratamento via TreeTagger e Spacy), buscas via CQPweb e apresentação de resultados de análises preliminares.

O estudo **"As contribuições da Linguística de Corpus e do Processamento de Linguagem Natural na elaboração do protótipo do Dicionário Ideológico de Locuções"**, de autoria de Thyago José da Cruz (CPTL/UFMS) inicialmente apresenta uma revisão bibliográfica teórica sobre a linguística de *corpus* e o processamento da linguagem natural. O autor demonstra o uso de recursos da linguística de *corpus* e de uma ferramenta computacional por meio do software *FieldWorks Language Explorer* (FLEx) para a prototipação do Dicionário Ideológico de Locuções, de caráter monolíngue, com locuções tanto onomasiológicas quanto semasiológicas. A extração das unidades fraseológicas para compor o dicionário adotou o *Corpus Brasileiro* e a *Web* como fontes primárias, e também o Tesouro do Léxico Patrimonial Galego-Português e o Dicionário de Expressões Idiomáticas como fontes secundárias. O artigo ainda descreve os passos executados no FLEx para a elaboração do corpo do dicionário, com a construção da parte sinóptico-analógica, da parte analógica e da parte alfabética.

O próximo artigo deste número temático, que fecha o primeiro eixo temático, **"Reflexões metodológicas sobre *datasets* e linguística de *corpus*: uma análise preliminar de dados legislativos"**, foi escrito por Lúcia de Almeida Ferrari (UFMG) e Evandro Landulfo Teixeira Paradela Cunha (UFMG). Inicialmente, os autores analisam a diferença entre a utilização de *corpora* e de *datasets* na linguística, apontando as potencialidades e limitações de cada um. Na sequência, demonstram as possibilidades de uso de um *dataset* para pesquisa linguística ao realizar uma análise quantitativa e exploratória de caracterização dos dados, além de uma análise lexical avaliando as mudanças ao longo do tempo em uma análise do tipo diacrônica. O estudo linguístico exploratório, parte de uma etapa de pré-processamento no *dataset*

da Base de Normas Jurídicas Brasileiras visando sua compreensão, o que apontou a existência de erros, trazendo a reflexão dos autores para a importância desta etapa principalmente para análises linguísticas quando os dados não contaram com a participação de um/a linguista em sua preparação. Em suas considerações finais, os autores trazem as principais observações acerca do *dataset* analisado e discutem tanto as questões metodológicas empregadas quanto o uso de ferramentas e métodos computacionais para a análise linguística diacrônica.

O segundo grupamento temático, estudos pautados pela linguística computacional, é aberto pelo sexto artigo, intitulado **"Bases lexicais verbais do português brasileiro",** de autoria de Roana Rodrigues (UFS), Marcella Lemos-Couto(UFSCar), Francimeire Leme Coelho(UFSCar), Isaac Souza de Miranda Junior(UFSCar) e Oto Vale(UFSCar). O estudo é uma contribuição sobre o estado da arte no que se refere às bases lexicais verbais do português brasileiro (PB). Os autores focam na descrição, sistematização e classificação dos verbos do PB com ênfase nos recursos descritivos que podem ser utilizados nas tarefas de PLN. Foram descritas e comparadas três bases de dados verbais do PB - VerbNet.Br, VerboWeb e Verbo-Brasil - com extensão superior a 1.000 lexemas verbais, com acesso disponível on-line e gratuitamente, e que sofreram atualização nos últimos 10 anos. O artigo traz uma análise crítica sobre as bases de dados verbais do PB, discutindo os seus pontos comuns e divergentes, e explora as informações sintático-semânticas presentes em cada base.

Em **"*Modeling the prosodic forms of Discourse Markers* (Para uma modelagem das formas prosódicas dos Marcadores Discursivos)"** os autores Tommaso Raso (UFMG), Albert Rilliard (Université Paris Saclay) e Saulo Mendes Santos (UFMG), motivados pela imprevisibilidade na determinação de um item lexical ser um Marcador Discursivo (MD), e pela falta de clareza na identificação da função específica de um item lexical, mesmo se já determinado como um MD, trazem uma dupla

contribuição. A primeira é uma proposta de nova solução linguística, baseada em parâmetros prosódicos, para a identificação dos MDs e as funções específicas desempenhadas pelos diferentes tipos de MDs. Para isso, os autores demonstram a natureza prosódica inerente às marcas formais dos MDs, distinguindo diferentes funções de natureza interacional veiculadas por eles. Como segunda contribuição, os autores apresentam a metodologia aplicada à descrição das diferentes pistas prosódicas e os passos adotados na modelagem computacional dos MDs, de forma a possibilitar a extração automática dos diferentes tipos de MDs a partir de novos dados, discutindo diferentes estratégias e os prós e contras de cada uma delas.

Em **"A construção de um banco de dados lexicográficos em XML a partir de dados dialetais: o Processamento Automático de Linguagem Natural (PLN)"**, Jorge Luiz Nunes dos Santos Junior(UFMS) e Aparecida Negri Isquerdo (UFMS) demonstram o uso de ferramentas informáticas voltadas para a criação e o gerenciamento de *corpora*, na tarefa de extração automática de dados. Para isso, partem de aspectos teóricos da Lexicografia, da Dialetologia e da Linguística Computacional para estabelecer parâmetros para a construção de uma base de dados em *XML* (*Extensible Markup Language)*. Adicionalmente, os autores demonstram os passos metodológicos que vêm sendo utilizados na coleta de dados em áudio do *corpus* dialetal do Projeto Atlas Linguístico do Brasil (ALiB). A estruturação dos dados resultante deste estudo permitirá a construção de uma aplicação web que servirá de suporte para a elaboração de um vocabulário dialetal *on-line*. Por fim, os autores refletem sobre a importância do planejamento da arquitetura usada para organizar e estruturar os dados provenientes da extração automática de um *corpus* dialetal.

Na sequência, o nono artigo deste volume, **"Modelação da valência verbal numa gramática computacional do português no formalismo HPSG"** de autoria de Leonel Figueiredo de Alencar (UFC-EMAp/FGV) e Alexandre Rademaker (EMAp/FGV), versa sobre a implementação da valência verbal na PorGram, uma nova

gramática computacional do português a partir da teoria gramatical formal lexicalista HPSG. Através de uma rica discussão demonstrativa de resultados comparativos entre análises sintáticas rasas e as análises sintáticas profundas, e diferentes modelos de análise, nota-se um ganho na adoção do modelo HPSG - um analisador sintático profundo - dada sua capacidade de integrar a descrição sintática e a descrição semântica num único nível de representação. A justificativa ao desenvolvimento da PorGram configura-se em ser esta uma alternativa de software livre e de código aberto em contraponto à LXGram. Os autores apresentam todo o arcabouço teórico-prático inerente ao desenvolvimento da primeira parte da PorGram, oferecendo seus principais indicadores com resultados satisfatórios.

Em seu artigo, **"PFN-PT: *A Framenet Annotator for Portuguese* (Anotação semântica automática: um novo Framenet para o português)"**, Eckhard Bick (*University of Southern Denmark*) apresenta um novo recurso *framenet* para a anotação semântica automática do português. Partindo de uma discussão da relação entre a complexidade linguística e do desempenho das ferramentas de anotação, nota-se que as decisões sobre o design do *framenet* dependem do idioma a que está vinculada a anotação, devido às variações dos seus aspectos sintáticos e semânticos. O *Parsing Framenet for Portuguese* (PFN-PT) configura-se como um recurso para a anotação semântica automática de Português capaz de delinear tal ponte sintático-semântica. O PFN-PT consiste em duas partes complementares, o *framenet* e o anotador de *frames*, ligados por uma representação de dependência semanticamente informada e orientada por valência fornecida por um analisador morfossintático. O artigo apresenta o processo de construção do PFN-PT abrangendo questões como tamanho, cobertura e granularidade do léxico, parâmetros diferenciadores de frames como valência, sintaxe e classe semântica, e por fim, são introduzidas as questões vinculadas à anotação automática de *frames*, que foca em *parsing* e *tagger* para *frames* e papéis semânticos baseados em regras. Rumo à conclusão do artigo, são apresentadas

estatísticas de distribuição de categorias e a avaliação doeste novo recurso computacional.

Abrindo o eixo temático final, processamento de língua natural, o artigo **"Manual de anotação como recurso de Processamento de Linguagem Natural: o modelo *Universal Dependencies* em língua portuguesa"** dos autores Magali Sanches Duran (UNESP), Maria das Graças Volpe Nunes (USP São Carlos), Lucelene Lopes (PUC-RS) e Thiago Alexandre Salgueiro Pardo (USP São Carlos) contribui para os estudos em PLN ao apresentar o manual de anotação fundamentado no modelo internacional *Universal Dependencies* (UD). Os autores ainda refletem brevemente sobre a necessidade de *corpora* anotados e sua aplicação como bases de treinamento para os modelos de Aprendizagem de Máquina voltados ao PLN. Na introdução os autores trazem um percurso sobre o papel do linguista ao longo da evolução do PLN, passando pela ascensão dos *corpora* anotados como recurso valioso para os novos métodos de PLN e terminando por discorrer sobre a importância dos manuais de anotação como parte indissociável de um esquema de anotação. Na sequência apresentam fundamentações teóricas sobre anotação sintática de *corpus* de língua portuguesa e sua relação com o desenvolvimento de *parsers* do português. Seguindo, eles apresentam o modelo UD, seu esquema de anotação, seus conjuntos de etiquetas morfossintáticas e relações sintáticas, trazendo as principais decisões tomadas na instanciação de suas diretrizes no português brasileiro. Por fim, exploram questões relacionadas ao desenvolvimento de manuais para a anotação de *corpora* em português brasileiro segundo o modelo internacional UD.

Concluindo o volume temático, o artigo de Jackson Wilke da Cruz Souza (UNIFAL-MG) e Ariani Di Felippo (UFSCar) intitulado **"*Evaluating a typology of signals for automatic detection of complementarity* (Avaliação de uma tipologia de sinais para a detecção automática da complementaridade)"** apresenta uma tarefa de validação da taxonomia de sinais (textuais) proposta anteriormente pelo primeiro

autor para a detecção automática das relações de complementaridade CST (*Cross-Document Structure Theory*) em um *corpus* multidocumental de notícias em português brasileiro. Inicialmente, os autores fornecem uma introdução ao CST, apresentando sua estrutura principal para a análise e uma visão geral da noção de complementaridade. Em seguida, apresentam o *corpus CSTNews* e a tipologia (ou taxonomia) de sinais usada na avaliação. A avaliação foi realizada utilizando-se algoritmos de diferentes paradigmas de Aprendizado de Máquina supervisionados disponíveis no software *Weka*. Seus resultados apontaram para um alto índice de acurácia geral (superior a 90%), indicando o potencial dos algoritmos usados na detecção automática das relações de complementaridade.

3 Palavras finais

Gostaríamos de agradecer ao editor-chefe da Revista Domínios de Lingu@gem, Guilherme Fromm, por ter aceitado a nossa proposta temática para compor o conjunto de temas em votação pelo conselho editorial da revista para o ano de 2022, bem como por todo o trabalho que efetuou para que esse volume se concretizasse.

Igualmente, agradecemos aos autores pelo interesse em compartilhar suas pesquisas e seus artigos, e ao corpo de pareceristas que avaliou e fez sugestões para o engrandecimento dos trabalhos submetidos.

Esperamos que esse volume ajude a divulgar os trabalhos sobre o tratamento computacional do PB em toda a sua riqueza temática e metodológica, e desperte o interesse e curiosidade dos leitores para investigar mais sobre o assunto e descobrir o vasto e promissor campo que se abre à nossa frente.

Referências Bibliográficas

BICK, E. **The Parsing System "Palavras":** Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. Aarhus, Denmark: Aarhus University Press, 2000.

CHURCH, K.; LIBERMAN, M. The future of computational linguistics: on beyond alchemy. **Frontiers in Artificial Intelligence**, v. 4, 2021. Disponível em: https://www.frontiersin.org/articles/10.3389/frai.2021.625341. Acesso em: 02 ago. 2022. DOI https://doi.org/10.3389/frai.2021.625341

FRANCIS, W. N.; KUČERA, H. **Computational Analysis of Present-Day American English**. Providence: Brown University Press, 1967.

FREITAS, M. C. **Linguística computacional**. São Paulo: Parábola, no prelo.

KAVINSKI, A. A corrida pelo processamento da linguagem natural. **MIT Technology Review**. 12 jan. 2012. Disponível em: https://mittechreview.com.br/a-corrida-pelo-processamento-de-linguagem-natura/. Acesso em: 02 ago. 2022.

OTHERO, G. A.; MENUZZI, S. M. **Linguística computacional:** teoria e prática. São Paulo: Parábola, 2005.

SVARTVIK, J. (org.). **The London Corpus of Spoken English:** Description and Research. Lund Studies in English 82. Lund: Lund University Press, 1990.

TURING, A. M. I.—Computing Machinery and Intelligence. ***Mind***, vol. 59, n. 236, p. 433–460, 1950. Disponível em: https://academic.oup.com/mind/article/LIX/236/433/986238. Acesso em: 02 ago. 2022. DOI https://doi.org/10.1093/mind/LIX.236.433