SIMPÓSIO TEMÁTICO 24:

Humanidades digitais: limites e possibilidades do digital nas fontes documentais de língua

Coordenadores: Jorge Viana Santos (UESB) e Patrício Nunes Barreiros (UEFS)

A automatização de edições de originais manuscritos com escrita não normatizada para corpora eletrônicos

Autores: Luiz Fernando Cardeal de Souza [1,5], Cristiane Namiuti Temponi [1,2,3,4]
Instituição: [1] UESB - Universidade Estadual do Sudoeste da Bahia, [2] CNPq - Conselho Nacional de Desenvolv. Científico e Tecnologico, [3] Fapesb - Fundação de Amparo à Pesquisa do Estado da Bahia, [4] Fapesp - Fundação de Amparo à Pesquisa do Estado de São Paulo, [5] IFBA - Instituto Federal da Bahia

Resumo: Para Namiuti, Santos e, Leite (2011), a garantia de fidedignidade às formas originais dos textos antigos é a base que qualquer estudo diacrônico deve pressupor. Conforme os autores, “A integração entre o tratamento filológico e o computacional na elaboração de corpus para o estudo do português brasileiro é especialmente importante para a preservação e divulgação do patrimônio histórico-linguístico [...]”. (NAMIUTI; SANTOS; LEITE, 2011, p.1). A tarefa de edição de textos antigos enfrenta muitas dificuldades, entre elas a dificuldade de acesso aos documentos, a fragilidade e raridade dos documentos e as complexidades envolvidas na transcrição e normalização da escrita nas edições que atendam as necessidades da Filologia e da Linguística. A edição de textos antigos para corpora eletrônicos pode se beneficiar das possibilidades da Linguística Computacional. Conforme Aluísio e Almeida (2006), os corpora eletrônicos são caracterizados por anotação e pode ocorrer em nível sintático, semântico, discursivo etc., o que pode ser feito manualmente, automaticamente ou semi-automaticamente, quando são feitas por softwares e complementadas por linguistas. Os softwares de anotação que existem para a edição especializada de textos antigos em formato eletrônico com controle de linguagem não são automáticos e não são programados para aprender as anotações feitas relativas aos níveis de edição, gerando aumento do esforço de trabalho dos linguistas, aumentando a possibilidade de erro. Para preencher essa lacuna, propomos o desenvolvimento de uma ferramenta que, acoplada a um software de edição especializada, o eDictor (Paixão de Sousa, Kepler e Faria, 2010), possa aprender as anotações de edição feitas: o eAssigner - software em desenvolvimento no Laboratório de Pesquisa em Linguística de Corpus (LAPELinC) como produto de pesquisa de mestrado realizada no Programa de Pós-Graduação em Linguística da Universidade Estadual do Sudoeste da Bahia (UESB, Vitória da Conquista) sob a orientação de Cristiane Namiuti (UESB) e Pablo Faria (UNICAMP).
Palavras-chave: edição automática, corpora eletronicos, linguistica computacional

A elaboração de um coletor e de um corpus de comentários extraídos de portais de notícias

Autores: Evandro Landulfo Teixeira Paradela Cunha [1,2], Gabriel Magno de Oliveira Silva [1], Virgilio Augusto Fernandes de Almeida [1,3]
Instituição: [1] UFMG - Universidade Federal de Minas Gerais, [2] LUCL - Leiden University Centre for Linguistics, [3] Harvard - Berkman Center for Internet & Society at Harvard University

Resumo: Um tipo de conteúdo muito relevante para pesquisadores em humanidades digitais - inclusive linguistas - são os comentários publicados por leitores de notícias em portais online. A análise desses comentários permite o estudo de questões linguísticas nos mais variados domínios, como o lexical, o morfossintático e o pragmático, além de uma série de outras questões em diversas áreas do conhecimento. Assim, torna-se necessário o desenvolvimento de ferramentas capazes de auxiliar a coletar e a organizar esse material. Nesse contexto, são apresentados aqui dois recursos: (a) um coletor de comentários de portais de notícias; e (b) um corpus composto por comentários de leitores do portal UOL. O coletor de comentários - desenvolvido em código aberto e livre para uso, modificação e distribuição - possui um funcionamento simples, permitindo que um usuário não-técnico colete todos os comentários de várias notícias após simplesmente informar uma lista de URLs provenientes dos principais portais de notícias brasileiros. No corpus de comentários do portal UOL, em fase final de coleta, pretende-se disponibilizar mais de um milhão de comentários extraídos a partir notícias publicadas entre 2012 e 2014 em várias seções do site. Aqui, são discutidos os resultados dos esforços para a elaboração desses dois recursos e apresentam-se, de maneira geral, os principais desafios e limitações conceituais e computacionais

enfrentados durante a sua elaboração. Além disso, é demonstrado o funcionamento de cada uma das ferramentas, de maneira a promover sua utilização entre os membros da comunidade.
Palavras-chave: coletor, corpus, portais de notícias

Abordagem computacional para a questão do acento no português brasileiro

Autores: Bruno Ferrari Guide [1]
Instituição: [1] USP - Universidade de São Paulo

Resumo: A presente comunicação tem como objetivo apresentar os resultados obtidos pelo projeto de mestrado defendido em agosto de 2016, intitulado "Abordagem computacional para a questão do acento no português brasileiro". O comportamento do acento possui algumas irregularidades interessantes que podem ser aproximadas através do uso de modelos probabilísticos. Para isso, o primeiro passo da pesquisa foi a compilação de um Corpus volumoso de dados idioma – O Corpus ABG.
A seguir, o foco da pesquisa foi desenvolver modelos probabilísticos que servissem de base para mecanismos de atribuição de acento. O primeiro modelo desenvolvido e testado pelo projeto foi baseado em N-gramas, que se vale das frequências extraídas de um corpus. Depois foi implementado o classificador bayesiano ingênuo que atribui o acento levando em conta diversas variáveis. As mesmas foram escolhidas durante o desenvolvimento da pesquisa e foram de fundamental importância para o desempenho do mesmo. Ambos os modelos puderam ser comparados também com implementações baseadas em propostas teóricas retiradas da literatura sobre o assunto, em específico, Bisol (1992) e Lee (1995). Todos os resultados então compõem o quadro de comparações quantitativas que formam a análise apresentada na pesquisa, que nos permite também fazer algumas indagações sobre propriedades do funcionamento do acento e sobre como seria mais eficiente modelar tal aspecto da língua.
Palavras-chave: linguística computacional, fonologia, probabilidade, prosódia

Constituição de corpus oral digital anotado: vantagens e desafios

Autores: Elisângela Gonçalves [1]
Instituição: [1] UESB - Universidade Estadual do Sudoeste da Bahia

Resumo: Uma preocupação dos pesquisadores de um projeto pioneiro em coleta e análise de corpus em português brasileiro, o NURC (Norma Urbana Culta), é a perda de um acervo tão importante que tem servido como base de análise para estudiosos em todo o Brasil, pois seus dados foram armazenados em fitas magnéticas, muitas das quais já foram destruídas pela ação do tempo. Ademais, as pesquisas com o NURC, por muito tempo, foram de acesso limitado, já que se encontram em livros impressos. Todavia, a realidade da pesquisa de corpus tem mudado com o desenvolvimento do áudio digital e a possibilidade de sua digitalização no computador e com a conversão do analógico para o digital, o que tem sido feito, inclusive, com o material do NURC, de modo a preservar esse acervo tão significativo. Inserido na perspectiva das Humanidades Digitais, o projeto Dados Orais da Microrregião de Vitória da Conquista, Bahia: Construção de um Corpus Oral Digital Anotado objetiva facilitar o trabalho de pesquisadores, disponibilizando-lhes dados digitais, o que lhes conferirá economia de tempo e um leque maior de dados sem ter que ir a campo para coletá-los, propondo-se, assim, a elaborar e implementar um corpus anotado em meio digital com dados orais dessa microrregião. Se a constituição desse tipo de corpus apresenta suas vantagens, não deixa de ter seus desafios, a iniciar pela decisão de como se fará a transcrição dos dados, se fonética ou grafemática; “manual” ou através de um programa de computador, como o ELAN, “um software [...] que permite a criação, edição, visualização e busca de anotações através de dados de vídeo e áudio” (CHRISTMANN et al, 2010). Se essa ferramenta pode agilizar o trabalho de transcrição de dados, será que dispensa o trabalho de revisor(es)? Essas são discussões a serem levantadas neste trabalho.
Palavras-chave: constituição de corpus, humanidades digitais, oralidade, Vitória da Conquista-Bahia

Linguística computacional, humanidades digitais e os desafios na mineração de um dicionário histórico-biográfico

Autores: Suemi Higuchi [2], Cláudia Freitas [1]
Instituição: [1] PUC-Rio - PUC-Rio , [2] FGV/ CPDOC - FGV/ CPDOC

Resumo: Uma das áreas de pesquisa mais férteis da Linguística Computacional está no desenvolvimento de ferramentas para extração automática de informações em grandes volumes de fontes textuais. A ideia é capacitar o computador a identificar entidades (como pessoas, lugares, instituições e eventos) e relações entre estas entidades através do processamento morfossintático e semântico de sentenças, permitindo aplicações como a extração de dados estruturados a partir dos textos, e o desenvolvimento de sistemas de perguntas e respostas. Mas para chegar lá um longo caminho precisa ser percorrido. A presente pesquisa apresentação busca traçar este caminho através da exploração de algumas ferramentas e técnicas de extração de informações tendo como objeto empírico o Dicionário Histórico-Biográfico Brasileiro (DHBB), obra concebida pelo Centro de Pesquisa e Documentação de História Contemporânea do Brasil (CPDOC) da Fundação Getulio Vargas (FGV), composto por cerca de 7.500 entradas de verbetes biográficos e temáticos relacionados ao cenário político nacional. O trabalho tem como objetivos principais apresentar tanto a etapa de pré-processamento do corpus, isto é, tarefas que envolvem conversão, limpeza, atribuição de metadados, normalização, etc, passando pela definição das categorias linguísticas a serem usadas para as entidades, até a etapa de extração das informações propriamente dita, e ainda em andamento. Refletiremos sobre as técnicas de reconhecimento de entidades mencionadas, a identificação das relações, o mapeamento dos padrões lexicais para a construção destas relações, a presença de fenômenos linguísticos nos textos (vaguezas, ambiguidades, anaforismos), e as possíveis estratégias para lidar com estes fenômenos. Sobre tudo isso, busca-se conhecer o estado da arte dos recursos léxicos existentes e disponíveis para o português.
Palavras-chave: extração de informações, entidades mencionadas, processamento de corpus

O 13 de maio de 1888: uma análise semântica

Autores: Ana Paula dos Reis Couto [1]
Instituição: [1] UESB - Universidade Estadual do Sudoeste da Bahia

Resumo: No presente trabalho, relacionado a pesquisa de mestrado em andamento no Programa de Pós-Graduação em Linguística (PPGLIN) da Universidade Estadual do Sudoeste da Bahia (UESB), orientado por Jorge Viana (UESB), objetivamos analisar o sentido do acontecimento do 13 de maio de 1888, dia em que foi sancionada a Lei 3353 (Lei Áurea), que aboliu a escravidão no Brasil. Partindo da análise de jornais do século XIX, levantamos a questão: *Como semanticamente se caracteriza a liberdade enunciada no acontecimento do 13 de maio de 1888?*. Embasados na Semântica do Acontecimento (cf. GUIMARÃES, 2002; 2011), por hipótese, postulamos que o 13 de Maio remete a duas palavras antagônicas: *liberdade* associada a *progresso* e *escravidão* associada a *atraso*, de modo que o funcionamento do sentido de liberdade, tal como expressada em documentos relacionados ao 13 de Maio, especificamente em jornais publicados no contexto da promulgação da Lei Áurea, relaciona-se a aspectos econômicos, políticos e jurídicos do país, reservando a segundo plano o escravo, enquanto pessoa, agora dita livre. Por recorte metodológico, tomando o periódico baiano *O Asteróide*, acessível pela plataforma da Hemeroteca Digital da Biblioteca Nacional, e usando o software de busca acoplada à plataforma, selecionamos enunciados que materializam o sentido de *liberdade* associado a *progresso*, bem como o de *escravidão* associado a *atraso*, em textos publicados em edições que circunscrevem o 13 de maio de 1888.
Palavras-chave: escravidão, liberdade, semântica do acontecimento

O objeto livro: a complexidade da forma e o digital

Autores: Jorge Viana Santos [1,2,3,4], Cristiane Namiuti Temponi [1,2,3,4]
Instituição: [1] UESB - Universidade Estadual do Sudoeste da Bahia, [2] CNPQ - Conselho Nacional de Desenv. Científico e Tecnológico, [3] FAPESB - Fundação de Amparo à Pesquisa do Estado da Bahia, [4] FAPESP - Fundação de Amparo à Pesquisa do Estado de São Paulo

Resumo: O grupo de pesquisadores do Laboratório de Pesquisa em Linguística de Corpus (LAPELINC) da Universidade Estadual do Sudoeste da Bahia (UESB), campus de Vitória da Conquista, têm desenvolvido e aplicado o método LAPELINC com um fluxo de trabalho que compreende três grandes etapas para a

construção de corpora eletrônicos anotados, cientificamente controlados (cf. Namiuti; Santos, 2015; Namiuti, Santos; Costa, 2016): (1) transposição; (2) transcrição; (3) compilação. Como demonstram Namiuti e Santos (2016), experiências como a do LAPELINC, na construção do corpus digital DOViC (Corpus de Documentos Oitocentistas de Vitória da Conquista), composto de manuscritos do século XIX, guardados nos arquivos do Fórum de Vitória da Conquista-Bahia (Brasil), tem servido para mostrar a importância, nem sempre lembrada, de que os resultados do trabalho de investigação, durante e após o desenvolvimento como um projeto, requerem ferramentas que atendam a necessidade de flexibilidade e automatização na recuperação de informação e reuso de tecnologias. Essa necessidade pode ser obtida com os sistemas de gerenciamento de informações, banco de dados e ferramentas de busca automáticas garantindo o fluxo de trabalho completo que envolve a integração de três objetos na construção de corpora digitais: Documento Físico (DF); Documento Digital Imagem (DDI); Documento Digital Texto (DDT). Ao considerar, na etapa (1), o uso da Fotografia enquanto um meio científico de transposição do texto em papel para o digital, Santos (2010a, 2010b, 2013) tem como objetivos centrais a reflexão sobre a complexidade do documento histórico a se tornar um documento digital imagem (DDI), que compreende o acesso, a forma, a fragilidade e/ou raridade, bem como a apresentação de técnicas de tratamento e fotografia destes documentos. No presente trabalho, objetivamos apresentar descritivamente o objeto livro digital, destacando soluções técnicas do Método LAPELINC para complexidades relativas à forma.
Palavras-chave: humanidades digitais, fotografia, livro

Sentido(s) de liberdade(s) encontrado(s) na primeira constituição republicana do Brasil

Autores: Anna Cláudia Pereira Queiroz [1]
Instituição: [1] UESB - Universidade Estadual do Sudoeste da Bahia

Resumo: Este trabalho, relacionado a pesquisa de mestrado em andamento no Programa de Pós-Graduação em Linguística (PPGLIN) da Universidade Estadual do Sudoeste da Bahia (UESB), orientado por Jorge Viana (UESB), investiga o(s) sentido(s) de liberdade encontrado(s) na redação da primeira constituição republicana brasileira, promulgada em 24 de fevereiro de 1891, mais especificamente, três anos após a abolição oficial da escravidão no Brasil, instituída pela Lei 3353 (Lei Áurea), sancionada em 13 de maio de 1888. Segundo Santos (2008), no período escravagista brasileiro, funcionaram dois sentidos de liberdade: uma *transitiva*, atribuída ao liberto, e outra *intransitiva*, esta característica dos senhores de escravos. Assim, questionamos: *Os sentidos de liberdade de Santos (2008) reaparecem na Constituição de 1891? E, se reaparecem, de que maneira? E, se não reaparecem, o que aparece no lugar?* Objetivamos dessa maneira verificar se, mesmo após a lei de abolição da escravidão, os sentidos apontados por Santos (2008), liberdade transitiva e liberdade intransitiva, permanecem funcionando no texto da primeira Constituição do período pós-escravagista. Em caso de registro deste reaparecimento, vislumbra-se demonstrar de que modo isso ocorreu; e, na hipótese de não ter acontecido, busca-se apontar qual(is) a(s) variação(ões) elencada(s) no referido diploma legal para o termo *liberdade*. Para tanto, recorre-se aos preceitos trazidos pela Semântica do Acontecimento (cf. GUIMARÃES, 2002; 2011) e, em determinados momentos, consultamos alguns pressupostos do Direito. A metodologia do trabalho define-se pela análise do *corpus* escolhido, em formato de texto digital pesquisável, que foi retirado da plataforma digital do site do Palácio do Planalto – Presidência da República *(www2.planalto.gov.br/).*
Palavras-chave: semântica do acontecimento, liberdade, direito

Texto ciborgue: reflexões epistemológicas sobre o texto como objeto da Filologia

Autores: Arivaldo Sacramento de Souza [1]
Instituição: [1] UFBA - Universidade Federal da Bahia, [2] UFBA - Universidade Federal da Bahia

Resumo: A renovação dos estudos filológicos produzidos a partir dos abalos dos preceitos da ciência positivista tem trazido questões que, cotidianamente, desafiam a ideia de um método universal para a crítica de texto. Noções como autoria, originalidade e texto têm imposto desafios de transformação das práticas de edição. No contexto das humanidades digitais, em que as transformações tecnológicas mudaram a relação do sujeito com o texto e a leitura, mas também com a experiência subjetiva com o corpo e a cidade, outras questões surgem, e alguns intelectuais estão atentos a isso, embora não discutam tais questões no labor dos filólogos. Em face do exposto, interessa-nos, aqui, uma reflexão metacrítica acerca do texto como objeto para os estudos filológicos no contexto da era digital. Para isso, construímos uma revisão sobre as concepções de texto que norteia a prática editorial teleológica, as que buscam o estabelecimento de texto, tal como a edição crítica. Em seguida, observamos como as prática de edições pragmáticas, aquelas que buscam apresentar o processo de criação, de produção, de circulação e de recepção, estão sendo

possíveis na construção do tecido digital. Com isso, visamos (i) perceber as transformações de concepções existentes para o texto enquanto objeto a partir das pesquisas de diversas fontes textuais; (ii) compreender que a concepção de unidade e de coerência textual é produto da crítica de texto, não princípio ontológico; (iii) refletir que o texto, próximo daquilo que Deleuze pensa sobre o "corpo sem órgãos", pode ser agenciado de modo a converter até o ato editorial em crítica (filológica, de certo). Neste último aspecto, tomamos as propostas de edição providenciadas por Eucanaã Ferraz na construção de cancioneiros de artistas da MPB, quais sejam Adriana Calcanhoto e Caetano Veloso.
Palavras-chave: texto digital, filologia, ciborgue

Topografia de documentos digitais imagem (DDI'S): um procedimento de indexação de informações do método LAPELINC

Autores: Jorge Viana Santos [2,3,4,5], Aline Silva Costa [1], Bruno Silvério Costa [1]
Instituição: [1] IFBA - Insituto Federal de Educação Ciência e Tecnologia da Bahia, [2] UESB - Universidade Estadual do Sudoeste da Bahia, [3] CNPq - Conselho Nacional de Pesquisas, [4] Fapesb - Fundação de Amparo ao Pesquisador do Estado da Bahia, [5] Fapesp - Fundação de Amparo à Pesquisa do Estado de São Paulo

Resumo: A intensificação do trabalho com textos antigos no âmbito da Linguística Histórica e de Corpus propiciou a busca de novas soluções técnicas para o gerenciamento, catalogação e edição especializada de textos antigos em meio eletrônico. Utilizando-se das vantagens da tecnologia, da computação e da linguística, o grupo de pesquisadores do Laboratório de Pesquisa em Linguística de Corpus (LAPELINC) da Universidade Estadual do Sudoeste da Bahia (UESB), campus de Vitória da Conquista, têm desenvolvido e aplicado o método LAPELINC com um fluxo de trabalho (workflow) que compreende três grandes etapas para a construção de corpora eletrônicos anotados, cientificamente controlados (cf. Namiuti; Santos, 2015; Namiuti, Santos; Costa, 2016): (i) transposição; (ii) transcrição; (iii) compilação. É nesse contexto que propomos a disponibilização dos textos antigos do corpus digital DOViC (Corpus de Documentos Oitocentistas de Vitória da Conquista), um corpus composto de manuscritos do século XIX, guardados nos arquivos do Fórum de Vitória da Conquista-Bahia (Brasil), por meio da implementação de um software baseado em tecnologias da Web, o WebSinc. Neste trabalho, por recorte, na etapa de transposição material do Documento Físico (DF) para o Documento Digital Imagem (DDI), focalizamos o sistema de banco de dados, gerenciado pelo aplicativo WebSinc, destacando o processo denominado topografia, por Santos e Namiuti (2016). O software desenvolvido contempla as atividades de gerenciamento para registrar topograficamente os DDI's através de interface gráfica, facilitando o trabalho dos administradores e compiladores do corpus em questão. Foram implementadas as funcionalidades de cadastro de documentos do corpus, fazendo o upload das imagens dos manuscritos e textos anotados.
Palavras-chave: linguística de corpus, banco de dados, aplicação web

Caderno de resumos do X Congresso Internacional da ABRALIN – Pesquisa linguística e compromisso político. / Organizadores: Anabel Medeiros de Azerêdo; Beatriz dos Santos Feres; Patrícia Ferreira Neves Ribeiro; Roberta Viegas Noronha; Silmara Dela Silva. Niterói: UFF, 2017.
Disponível em: <http://abralin.org/congresso2017/programacao-1?prog=simposios>.