Revista RBBA

Revista Binacional Brasil Argentina

COMPUTAÇÃO E LINGUÍSTICA: IMPORTANTE DIÁLOGO PARA PESQUISAS E PRESERVAÇÃO DA MEMÓRIA NOS NOVOS MEIOS DAS ANTIGAS FONTES

COMPUTACIÓN Y LINGÜÍSTICA: IMPORTANTE DIÁLOGO PARA INVESTIGACIONES Y PRESERVACIÓN DE LA MEMORIA EN LOS NUEVOS MEDIOS DE LAS ANTIGUAS FUENTES

Cristiane Namiuti Temponi
Jorge Viana Santos
Aline Silva Costa
Igor Sodré Farias
Universidade Estadual do Sudoeste da Bahia (UESB)/ cristianenamiuti@pq.cnpq.br
Universidade Estadual do Sudoeste da Bahia (UESB)/ viana.jorge.viana@gmail.com
Universidade Estadual do Sudoeste da Bahia (UESB)/ aline@uesb.edu.br
Universidade Estadual do Sudoeste da Bahia (UESB)/igsfarias@gmail.com

Resumo
Neste artigo, exploramos alguns desafios atuais da pesquisa em Lingüística de Corpus, na sua vertente dedicada à História das Línguas – focalizando, particularmente, a experiência do trabalho com textos históricos da Língua Portuguesa em meio eletrônico dando notícias do trabalho que vem sendo desenvolvido na Universidade Estadual do Sudoeste da Bahia, no âmbito dos projetos: (i) *Corpora* Digitais Para a História do Português Brasileiro - região Sudoeste da Bahia: Aliança PHPB – Tycho Brahe (FAPESB); (ii) Novos meios para antigas fontes: sintaxe diacrônica em *corpus* eletrônico (UESB/FAPESB); e (iii) O Português no tempo e no espaço (FAPESP). Sustentaremos alguns caminhos que já se mostram promissores na exploração da fronteira da pesquisa representada pela união da Lingüística e da Computação.
Palavras-chave: Documentos Antigos. Fotografia. Computação. Linguística.

Resumen
En este artículo, exploramos algunos de los retos actuales de la investigación en Lingüística de Corpus en su capítulo sobre la Historia de las Lenguas - centrándose, en particular, la experiencia de trabajar con textos históricos de la lengua portuguesa en formato electrónico, dando noticia de la labor que se está desarrollado en la Universidade Estadual do Sudoeste de Bahia, en el marco de los proyectos: (i) *Corpora* Digitales para la Historia del Portugués Brasileño – región Suroeste de Bahía: Alianza PHPB – Tycho Brahe (FAPESB); (ii) Nuevos medios para antiguas fuentes: sintaxis diacrónica en corpus electrónico (UESB / FAPESB); y (iii) El portugués en el tiempo y en el espacio (FAPESP). Mantendremos algunos caminos que ya parecen ser prometedores en la exploración de la frontera de la investigación, representada por la unión de la Lingüística y de la Computación.
Palabras-clave: Documentos antiguos. Fotografía. Computación. Lingüística.

Introdução

A investigação diacrônica depende dos textos antigos. No entanto, apesar dessa necessidade premente, no Brasil, grande parte dos documentos históricos que sobreviveram ao tempo, seja em arquivos públicos, privados ou pessoais, não está acessível do ponto de vista científico (enquanto *corpus* manipulável) ao pesquisador, nem tampouco do ponto de vista material ao cidadão. Apesar da dificuldade para acessar o dado que interessa aos estudos da linguagem realizados sob uma perspectiva histórica – seja na diacronia ou na sincronia –, muito tem sido feito desde as propostas de pesquisa anunciadas no final da década de 1990 por Galves, Castilho e Mattos e Silva, idealizadores e coordenadores dos seguintes projetos de pesquisa: (i) Padrões Rítmicos Fixação de Parâmetros e Mudança Lingüística (GALVES et al. 1997); (ii) Para a História do Português Brasileiro (CASTILHO, 1997); (iii) Programa para a História da Língua Portuguesa (MATOS E SILVA, 1992).

O trabalho que vem sendo desenvolvido na Universidade Estadual do Sudoeste da Bahia, pelo grupo do Laboratório de Pesquisa em Lingüística de Corpus (Lapelinc) no âmbito dos projetos parceiros - (i) Corpora Digitais Para a História do Português Brasileiro - região Sudoeste da Bahia: Aliança PHPB – Tycho Brahe (FAPESB: 6171/2010) (SANTOS;

NAMIUTI, 2010), (ii) O português no tempo e no espaço (FAPESP: 2012/06078-9) (GALVES, 2012) (iii) Novos meios para antigas fontes: sintaxe diacrônica em corpus eletrônico (UESB) (NAMIUTI, 2010), (iv) Sintaxe diacrônica em corpus eletrônico: do português pré-clássico às variantes modernas (Convênio FAPESB/UESB 006/2012) (NAMIUTI, 2012) e (v) Memória da escravidão baiana: análise semântica comparativa de sentidos de liberdade em cartas de alforria oitocentistas de Vitória da Conquista e Rio de Contas (Convênio FAPESB/UESB 006/2012) (SANTOS, 2012) - visa contribuir com o trabalho com textos antigos no Brasil.

De acordo com Paixão de Sousa (2006), a questão central que se coloca para o trabalho com textos antigos como fundamentos para estudos lingüísticos no meio eletrônico é a busca por uma abordagem global do texto, em termos conceituais e tecnológicos, que se reflita numa integração entre diferentes planos de análise. De fato, os estudos históricos realizados com base em textos antigos dependem, antes de tudo, da garantia da fidelidade às formas originais dos textos – sendo este o pilar de sustentação que qualquer estudo lingüístico, em qualquer quadro teórico, deve pressupor. No caso dos *corpora* eletrônicos, esse pressuposto fundamental precisa ser integrado com requerimentos impostos pela vertente computacional e lingüística dos estudos – tais sejam: o arquivo virtual/digital, a confiabilidade e durabilidade do código, a necessidade de quantidade, agilidade e automação no trabalho de organização e seleção de dados.

Namiuti, Santos e Leite (2011) enfatizam que a integração entre o tratamento filológico e o computacional na elaboração de *corpus* para o estudo do português é especialmente importante para a preservação e divulgação do patrimônio histórico-linguístico do sudoeste baiano, e, por extensão, da Bahia e do Brasil. Nessa linha, buscamos abordar a importância da Lingüística de *Corpus* e dos Estudos Diacrônicos para a preservação da memória (linguística e cultural), apresentando os desafios metodológicos impostos pela ciência Lingüística aliada às vertentes tecnológica (Fotografia) e computacional (sistemas de gerenciamento de informação, banco de dados, edição e anotação eletrônicas de textos), bem como alguns desenvolvimentos metodológicos envolvendo essas duas vertentes combinadas para o trabalho com textos antigos.

Do papel ao texto digital: notas sobre a digitalização e edição de documentos manuscritos e impressos

Paixão de Sousa (2004, 2006) mostrou que a combinação de controle e flexibilidade no processamento dos textos garante o rigor da descrição lingüística, ao mesmo tempo em que possibilita a multiplicação do alcance do corpus para diferentes usuários finais. De acordo com Paixão de Souza (2004,2006) a potencialização do alcance de uso dos textos reforça seu valor como fonte de pesquisa para diferentes áreas do conhecimento, e colabora com a preservação e divulgação de um patrimônio histórico e cultural livremente disponível via rede mundial de computadores. Lembra ainda que a otimização dos processamentos linguísticos consolida a finalidade principal do corpus, que é fornecer com agilidade e precisão um grande volume de dados para a análise lingüística.

A busca por uma abordagem global do texto, em termos conceituais e tecnológicos, que se reflete numa integração entre diferentes planos de análise, constitui, segundo Paixão de Sousa (2006), uma questão central que se coloca para o trabalho com textos antigos como fundamentos para estudos lingüísticos no meio eletrônico. Para a autora, "um corpus histórico eletrônico pode ser concebido como um conjunto de textos escritos em épocas passadas e reunidos em torno de uma determinada concepção de língua, com o objetivo fundamental de constituir um corpo robusto e tecnologicamente trabalhável de informações que possibilitem análises lingüísticas aprofundadas". E ainda destaca que, o material componente desse corpo de informações percorre um longo caminho até sua transformação em arquivos computacionalmente estruturados, em cujo percurso, "(...) os diferentes estágios de processamento imprimem modificações que podem constituir informações importantes para a compreensão do significado histórico e lingüístico dos textos". Assim, Paixão de Sousa (2006) argumenta que os rastros que são deixados nos documentos nesse caminho entre o momento em que são escritos e sua instrumentação computacional constituem o que denomina de memórias do texto.

Enfim, com base nesses fundamentos, podemos afirmar que essas memórias, ao serem capturadas e codificadas tecnologicamente, contribuem para a construção de conhecimentos de ordens diversas, provenientes do legado desses textos.

Ao considerar o uso da Fotografia enquanto um meio científico de transposição do texto em papel para o digital, Santos (2010a, 2010b, 2013) procura enfatizar a necessidade de

se colocar na posição de um Pesquisador Formador de Corpora (PFC) e não apenas de um Pesquisador pragmático, e, tem como objetivos centrais a reflexão sobre a complexidade do documento histórico a se tornar imagem digital - o acesso, a forma, a fragilidade e/ou raridade, bem como a apresentação de técnicas de tratamento e fotografia destes documentos. O trabalho de Santos perpassa as seguintes questões centrais: "Qual a viabilidade do uso da fotografia para a captação fidedigna de documentos para compor corpora digitais, visando estudos linguísticos e científicos?", e "Quais são as complexidades intervenientes no processo de digitalização de documentos físicos escritos?", defendendo a hipótese de que, desde que metodicamente controlada em suas fases de captura, catalogação, edição, armazenamento, e leitura, a Fotografia apresenta-se como forma altamente viável e producente de digitalização, permitindo à Linguística, ou outra ciência, acessar imageticamente, de modo confiável, o documento não disponível no local da pesquisa.

Nesse sentido, no tocante à complexidade do documento histórico, Santos (2010a, 2010b) destaca em primeiro lugar o acesso ao local em que se encontram. Afirma que, por estarem normalmente em arquivos institucionais, e às vezes pessoais, nem sempre se tem acesso franqueado a eles por longo tempo ou por repetidas vezes. Assim, quando se consegue acesso a este tipo de arquivo, é necessário colher o maior número de informações documentais possível, pois a informação (tecnicamente, o metadado) que pode não interessar imediatamente a um pesquisador, poderá ser fundamental para outro. Um exemplo, em Linguística, é a história que cerca o próprio documento arquivado.

Em segundo, considerando a forma, o autor enfatiza que, sendo físico, tridimensional, não padronizado, como é o caso de manuscritos históricos, e apresentando-se numa organização que, necessariamente, precisará ser desfeita (ou melhor: virtualizada) na etapa de digitalização fotográfica, um livro, por exemplo, que tem páginas (às vezes sem identificação específica) todas presas e por isso automaticamente identificadas, será fotografado página por página, nem sempre na mesma ordem. Assim, registrar essa ordem antes, durante a Fotografia e na própria imagem (fotografia) é, pois, fundamental, visto que a digitalização pode tornar-se improducente caso não seja acompanhada da devida identificação (recuperável) da organização original dos textos.

E em terceiro, quanto à fragilidade e/ou raridade dos documentos, Santos (2010a, 2010b) destaca que ela impõe ao PFC muito cuidado no manuseio para evitar danos à fonte, fato que requer, como condição *sine qua non* o conhecimento e uso de certas técnicas e

equipamentos que, ao mesmo tempo em que garante um registro fidedigno, minimizam a possibilidade de dano ao original.

Considerar esses três aspectos em conjunto no uso da Fotografia cientificamente controlada é uma das características de um PFC. Nesse sentido, para o uso da Fotografia científica nas práticas do Lapelinc (Laboratório de Pesquisa em Linguística de Corpus/UESB), vimos desenvolvendo e aplicando um método específico de fotografia (cf. Santos 2013), que, dentre outros aspectos técnicos, pressupõe que a fotografia não é num mero meio de reprodução de um documento, como vemos no gráfico 1, ou seja, uma fotografia pragmática, que serve apenas a uma pesquisa e não tem compromisso de futuro: a fotografia tal como praticada por um pesquisador pragmático:

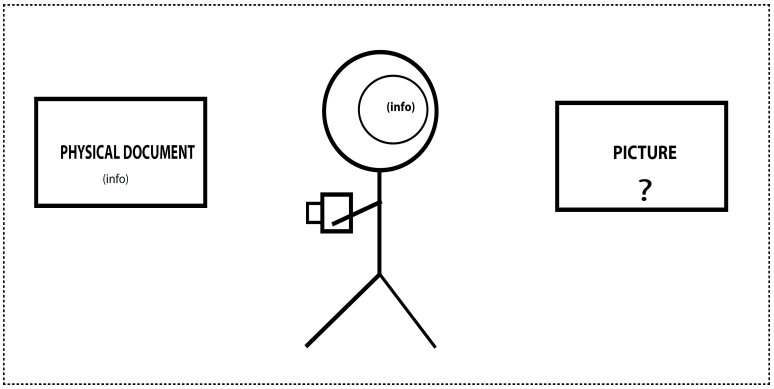

GRÁFICO 1: Fotografia como meio simples de reprodução digital por um pesquisador pragmático[i].
Fonte: Elaboração própria

Como se vê, este tipo de fotografia, sempre foi e será praticado: trata-se da foto feita com um simples compromisso de acessar o documento através da leitura da imagem: não requer maiores conhecimentos. É a foto que um pesquisador faz para si, com finalidade, por exemplo, de copiar documentos para uma tese. Em suma, não é feita para corpus, mas para uso.

Mas a situação é muito diferente se olharmos para o gráfico 2:

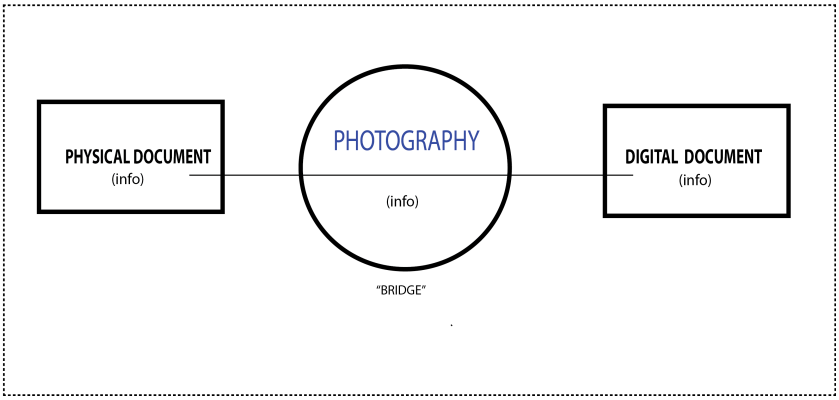

GRÁFICO 2: Fotografia praticada com método científico de reprodução digital: a ponte entre Documento físico e Documento digital.[ii].
Fonte: Elaboração própria

No gráfico 2, observa-se que a Fotografia funciona como uma espécie de *ponte* entre o DF (documento físico) e o DD (documento digital). Mas para isso, defendemos a necessidade de que ela registre, na própria imagem, dados/informações que façam com que a imagem gerada não perca o vínculo com o documento que lhe deu origem.

Como fazer isso? Como sustentar que a fotografia (*photograph, picture*) que resultará do processo (*photography*) será fidedigna e continuará vinculada ao documento original? Com um método específico, a exemplo do que utilizamos: O *Método Lapelinc*, proposto por Santos (2013), que dentre outros aspectos, possui etapas como: Controle e Captura de informações da fonte); Captura fotográfica dupla da imagem do original; Catalogação no Database DOViC (Documentos oitocentistas de Vitória da Conquista) das imagens componentes do documento (por exemplo, capa, folhas de um livro...); Gerenciamento de etapas de construção do corpus, com o SGP (Sistema de Gerenciamento de Pesquisa).

Enfim, ressalte-se que, assim praticada, no interior de um método, a Fotografia pode ser cientificamente controlada, trazendo, sem dúvida, vantagens para as mais diversas pesquisas, tais como: a) uma nova forma de acesso da imagem: a visual, que, ao mesmo tempo em que, sendo digital, permite sua veiculação virtual, pode com isso democratizar a consulta de fontes que, sendo únicas (e em papel), só podem ser vistas/consultadas *in locu*; b) a manipulação visual eletrônica do texto (ampliação, contraste, cor...); e – muito importante – a reprodutibilidade: um documento adequadamente digitalizado com auxílio da Fotografia metodologicamente controlada, tal como fazemos no Lapelinc, pode ser reproduzido eletronicamente, o que não só contribui para a preservação do original, como também

possibilita sua consulta e análise por pesquisadores de áreas as mais diversas. Ou seja: ao mesmo tempo, a Fotografia viabiliza a ciência, e preserva a história, a nossa memória.

O documento digital capturado pelo Método Lapelinc gera nosso documento original para a transcrição, edição e anotação nos mesmos moldes do *Corpus Histórico do Português Tycho Brahe*[iii], com o auxílio da ferramenta *eDictor* (PAIXÃO DE SOUSA; KEPLER; FARIA, 2010). O texto transcrito é salvo em um arquivo no formato texto simples (TXT). Edições como modernização, junção, segmentação e modernização de grafia são feitas por meio da interface gráfica da ferramenta, produzindo como resultado um arquivo anotado na Linguagem de Marcação Extensível (*eXtensible Markup Language* - ***XML***, cf. W3C, 1999). O *eDictor* permite que a edição dos textos seja realizada sem a necessidade do contato direto dos editores com a linguagem XML. Como argumentam os autores o uso dessa ferramenta trouxe diversos benefícios à tarefa de edição, sendo os principais a diminuição significativa nas revisões (que ficaram restritas à revisão dos aspectos linguístico-filológicos) e a garantia de uma marcação XML bem-formada". O software faz também a anotação das informações morfológicas dos textos, usando o mesmo formato XML, e gerando um arquivo único com as várias versões do texto: o texto transcrito tal qual o original, as edições (filológicas, modernizadas e/ou técnicas) e ainda a anotação morfológica das palavras no texto) .

Possibilidades de processamento automático nos novos meios das antigas fontes

Como aponta Paixão de Sousa (2007), na etapa de transcrição dos textos, já se destacam para o processo de trabalho as singularidades técnicas do meio eletrônico. Ao se transcrever ou digitalizar um texto – ou seja: na passagem do meio físico para o meio digital – está-se alterando substantivamente o sistema de codificação da informação, de visual para computacional-matemático. Esta passagem, se realizada de forma não-sistemática, encerra grande potencial de perda de informações, em detrimento da fidedignidade ao texto original. Em textos manuscritos ou impressos, a seqüência de caracteres que forma o texto, bem como diversas informações estruturais importantes (por exemplo, a paragrafação), é codificada de modo direto e visual. Em textos processados eletronicamente, essas informações são codificadas indiretamente por programas de processamento de texto.

Paixão de Sousa (2004) recomenda que a produção de textos em meio eletrônico com finalidades específicas (por exemplo - construção de corpora de língua), se deve fazer uso de

um processamento controlado que permita a codificação de uma grande variedade de informações, de modo confiável e transportável.

Seguindo este pensamento, no processamento eletrônico de textos, as estruturas precisam ser anotadas em alguma linguagem de anotação, e depois "traduzidas" ou "lidas" por uma programação que gera a apresentação final do texto . No Corpus Tycho Brahe usa-se a linguagem de anotação XML.

Os documentos são anotados conforme as diretrizes e os procedimentos seguidos no sistema de edição:

(i) a catalogação dos textos;
(ii) a transcrição dos textos;
(iii) a codificação da interferência editorial sobre os textos;
(iv) a apresentação dos textos

Após a etapa de transcrição e anotação das estruturas gráficas, os textos passam pelo processo de edição (tomando aqui o termo “edição” no sentido mais estrito, de interferência interpretativa em relação ao texto original). Incluem-se, neste plano, diferentes graus de interferências de edição – das mais restritas, próprias das edições paleográficas (desdobramento de abreviaturas; decisões de leitura), às mais amplas, próprias das edições modernizadas (atualização de grafia).

Em termos substantivos, seguem-se as normas estabelecidas para as edições filológicas em geral. Entretanto, neste âmbito das interferências no texto há uma singularidade crucial do trabalho de edição eletrônica: este sistema explora as possibilidades próprias do suporte informático de modo a permitir a manutenção do texto original no mesmo plano em que se realizam as interferências editoriais. Assim, o documento eletrônico usado pelo editor contém todas as informações de transcrição e de edições, devidamente codificadas, de forma a garantir a integridade do texto transcrito do início ao fim do processo. Dito de outra maneira, as palavras (e todo o texto nas suas respectivas versões e graus de interferências) são mapeadas, e, por isso, podemos transitar pelas edições e recuperar as informações da palavra original no texto modernizado (eg. Corpus Tycho Brahe). É esta a característica que confere controle e confiabilidade às edições eletrônicas assim desenvolvidas. Além de codificar as estruturas de texto (parágrafo, sentenças), a linguagem XML permite adicionar informações

ao texto transcrito, e anotá-las da mesma forma. Consideram-se, fundamentalmente, nos textos do Corpus Tycho Brahe, informações de três tipos:

a) Meta-informações: informações sobre o texto (autor, data, créditos, fonte, etc.)
b) Comentários do editor: características do texto original, decisões de leitura, etc.
c) Modificações: no caso de edições interpretativas, por exemplo, itens com grafia modernizada.

Tomando por base a estrutura básica dos textos, pode-se inserir em seguida informações adicionais, como:

(1) Comentários do editor - decisões de leitura. Anotação:
<ed_mark>comentário (exemplo - "borrado") </ed_mark>

(2) Expansão de abreviaturas. Anotação:
<v><ed> termo editado </ed><or> termo original </or></v>

Isso resulta em um documento que será transformado, na etapa seguinte, em uma versão do tipo edição semi-diplomática. Em seguida, adiciona-se mais uma camada de edição:

(3) Modernização de grafias. Anotação:
<v><ed_2>termo editado </ed_2><or>termo original </or></v>

Após essa etapa, obtém-se um documento a partir do qual, além da apresentação semi-diplomática, pode ser gerada uma apresentação modernizada. Esta anotação XML foi feita manualmente, pela equipe do corpus Tycho Brahe, utilizando o editor de texto simples – EMACS – por apresentar a anotação de comandos em cores diferentes, o que facilita a visualização da estrutura e do texto.

Com o objetivo de facilitar esse trabalho da anotação da estrutura XML, no ano de 2007 foi idealizada por Maria Clara Paixão de Souza (USP) uma ferramenta computacional – *eDictor*, que mais tarde foi construída e implementada com o auxílio de Fábio Kepler (UNIPAMPA) e Pablo Faria (PG-UNICAMP), especialistas da área de computação. O

EDictor é destinado à transcrição e codificação de textos em formato XML, para sua posterior edição e ao seu uso diverso, por exemplo, em análises linguísticas (morfológica, sintática, entre outras). O padrão XML adotado para a ferramenta foi pensado para abarcar informações de edição e de etiquetagem (morfológica).

O último passo do processo de preparação é a transformação dos documentos XML nos documentos finais – ou seja, a geração de versões a partir da anotação básica. A transformação é feita por meio de um programa capaz de ler as estruturas anotadas em XML e trabalhá-las conforme desejado. No Corpus Tycho Brahe usa-se um programa aberto e gratuito: o Saxon. Para configurar as transformações, é preciso utilizar a linguagem de programação XSLT. Sua sintaxe básica segue os mesmos princípios do XML (por exemplo, quanto ao encaixamento); as programações se ativam a partir da leitura da árvore (chamada *X-path,* o "caminho X") formada no documento-base. Cada programação corresponde a uma Folha de Estilos para Transformação (um documento *.xsl*).

Os comandos de programação XSLT permitem:

a) Acrescentar estruturas ao documento XML original como, por exemplo, informações repetitivas que seriam trabalhosas de adicionar manualmente.

b) Selecionar nódulos da árvore do XML original. Por exemplo, pode-se "chamar" apenas as estruturas de cabeçalho, gerando um documento de ficha catalográfica, sem o corpo do texto.

c) Reordenar nódulos da árvore do XML original. Por exemplo, pode-se "chamar" as estruturas de cabeçalho depois das estruturas do corpo de texto, gerando um documento "inverso" ao documento de base.

d) Formatar elementos do documento XML original, acrescentando estruturas de HTML clássicas (tamanho de fonte, cores, etc.).

A transformação é, portanto, a etapa na qual se obtém a apresentação final, como resultado de todo o processo de anotação.

Este sistema se fundamenta no processo integral e controlado de interferências editoriais. A partir daquele documento "de base", no qual o editor registrou, controladamente, todo o processo, da transcrição à modernização da grafia, é possível "extrair" diferentes formas de apresentação final do texto, sem que para isto seja necessário realizar qualquer

alteração no documento integral. Esta extração ou geração de versões para apresentação é realizada com grande agilidade, através de uma programação computacional simples; o processo pode, assim, ser repetido quantas vezes for necessário (o que permite, fundamentalmente, que se absorvam eventuais alterações ou correções na transcrição ou na edição do original).

Ferramentas de anotação e busca

Amparados nas soluções técnicas para a edição especializada de textos antigos em meio eletrônico propomos uma reflexão sobre o trabalho com dado de língua nesse meio, especialmente no que se refere às possibilidades de investigação e busca de dados propiciadas pela criação e implementação de analisadores automáticos empregados nos textos do Corpus Tycho Brahe: o *tagger* integrado à ferramenta Edictor (FARIA, KEPLER e PAIXÃO DE SOUZA, 2010) e o *parser* - ferramenta de anotação sintática de corpora desenvolvido pela Universidade da Pensilvânia - O *Penn TreeBank Forma*t (Formato *Penn TreeBank*) (SANTORINI, 2010; MARCUS;TAYLOR, 2002). Nesse *corpus,* o dado de língua está disponível em três formatos para a pesquisa:

(i) texto ortograficamente transcrito: Senhor: *Ofereço a Vossa Majestade as Reflexões sobre a vaidade dos homens*

(ii) texto morfologicamente etiquetados: *Senhor/NPR :/. Ofereço/VB-P a/P Vossa/PRO$-F Majestade/NPR as/D-F-Reflexões/NPR-P sobre/P a/D-F vaidade/N dos/P+D-P homens/N-P*

(iii) texto sintaticamente anotado:
```
(IP-MAT (NP-SBJ *pro*)
        (NP-VOC (NPR Senhor))
                (. :)
        (VB-P Ofereço)
        (PP (P a)
            (NP (PRO$-F Vossa) (NPR Majestade)))
        (NP-ACC (D-F-P as) (NPR-P Reflexões)
                           (PP (P sobre)
                               (NP (D-F a) (N vaidade)
                                   (PP (P+D-P dos)
                                       (NP (N-P homens))))))
```

O sistema de anotação sintática é uma árvore que pode ser visualizada sob a forma de parênteses etiquetados, como no exemplo acima, ou no formato arbóreo quando visualizada no Corpus Draw (cf. figura 1).

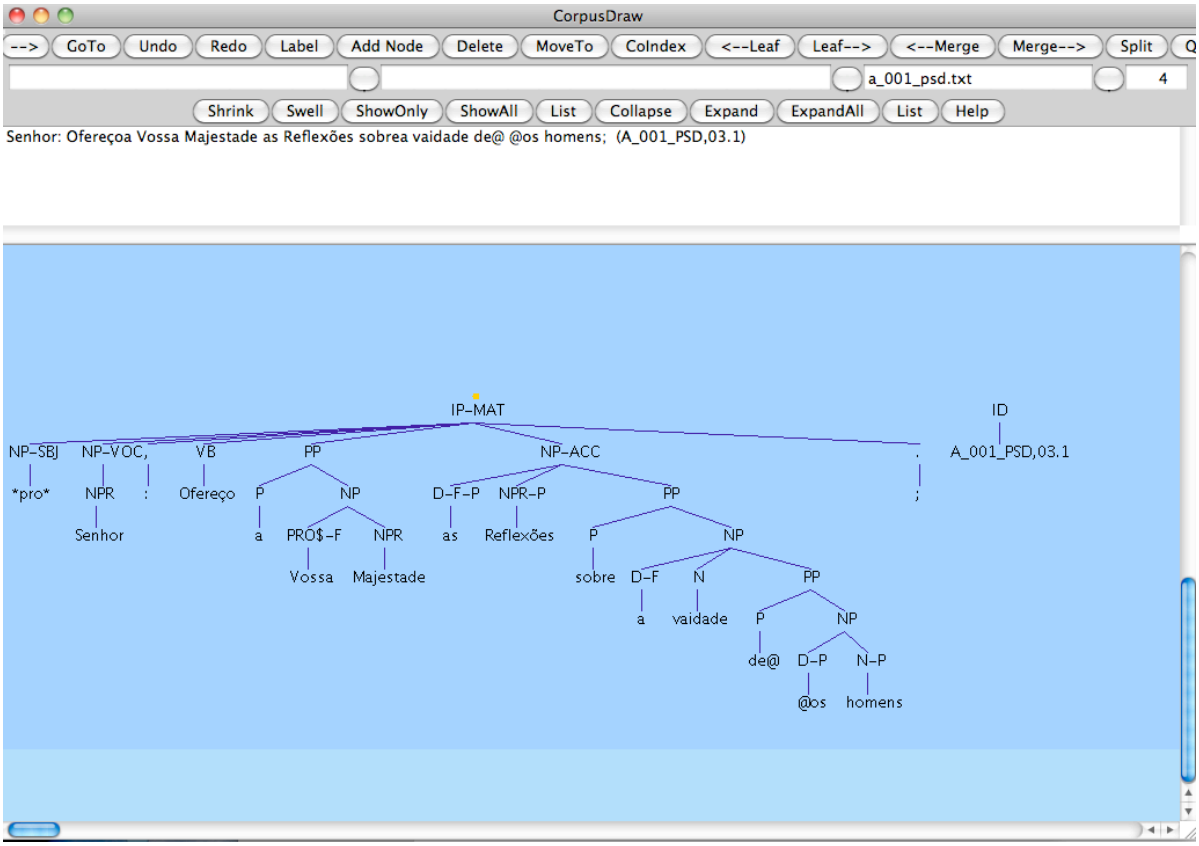

Figura 1: Representação em árvore da anotação sintática da sentença.

Fonte: Elaboração própria (*print-screen* da tela do programa)

O texto sintaticamente anotado é o formato que agrega o maior contingente de informações e, portanto é o formato com maior possibilidade de busca. Com este formato podemos recuperar desde as informações de ordem lexical até as informações de cunho gramatical - referentes a função sintática. Porém, atualmente, não há um anotador (*parser*) sintático acoplado a ferramenta *eDictor*. Isto implica em manter dois arquivos para um mesmo texto do corpus - um arquivo XML, guardando todas as informações, edições e anotação morfológica, e outro arquivo em formato TXT guardando apenas as informações da anotação sintática.

A busca automática de dados de língua pode ser realizada por meio de criação de *expressões regulares* e/ou *queries,* utilizando–se de linguagens de processamento de dados como por exemplo *perl* e de programas a exemplo de *Corpus-Search*.

O Corpus Search é um programa que realiza pesquisas sintáticas em corpora anotados no formato *Penn TreeBank*. Asssim como o esquema de anotação, o software também foi desenvolvido na Universidade da Pensilvânia (CORPUS SEARCH, 2009).

Implementado na linguagem de programação Java, *Corpus Search* é, portanto, multiplataforma e requer que o programa JRE (*Java Runtime Environment*) esteja instalado no computador do usuário.

A execução de *Corpus Search* para realizar buscas sintáticas requer duas entradas: o arquivo do corpus, anotado no formato *Penn TreeBank*; e o arquivo com a especificação da consulta a ser realizada, também chamado de *command file*, em formato de texto simples com extensão ".q", pois, esse arquivo deve conter, além dos comandos de controle de busca (obrigatório - *Node* - e opcionais) e especificações de saída (opcional), a expressão (algoritimo) de especificação da estrutura a ser pesquisada (*Query*).

A especificação das buscas no arquivo de entrada deve estar de acordo com a sintaxe exigida pela linguagem de consulta, que compreende chamada a funções de busca e uso de operações lógicas. As funções de busca pesquisam relações existentes na estrutura sintática como dominância, c-comando[iv], irmandade, entre outras. A ferramenta possui uma linguagem simples e mais fácil de aprender. Como o *Corpus Search* é específico para buscas sintáticas, os comandos foram projetados para este fim, apesar de possibilitar outros tipos de busca – por itens lexicais ou categorias morfológicas, ou busca mais sofisticadas – itens lexicais ou categorias morfológicas no interior de uma construção sintática específica.

```
define: port.def
print_indices: t
node: IP*
query: (IP-SUB iDoms NP-ACC)
AND (IP-SUB iDoms tns_vb)
AND (NP-ACC iPrecedes tns_vb)
```

Quadro 1: exemplo de *command file*

Fonte: Elaboração própria

A busca exemplificada no quadro 1 verifica a existência da expressão prevista na *query* em todo nó IP[v] de um arquivo de entrada, imprimindo índices nos dados encontrados. A *query* apresentada no quadro busca as orações subordinadas (IP-SUB) que dominam imediatamente (comando iDoms) um objeto direto (NP-ACC) e um verbo finito (tns_VB). O objeto direto precede imediatamente (comando iPrecedes) o verbo finito.

Os resultados de uma busca realizada pelo *Corpus Search* podem ser vistos no arquivo de saída (*output*) gerado pelo programa. O arquivo é produzido no formato texto simples e reúne informações sobre as sentenças contendo as restrições especificadas pela busca (CORPUS SEARCH, 2009).

O *output* contém um cabeçalho com informações sobre a data da busca, o arquivo de entrada (*input*), ferramenta utilizada e a expressão da busca (*query)*. Os dados são impressos logo após o cabeçalho da seguinte forma:

1) a sentença sem anotação

2) informação sobre o(s) dado(s) considerado(s) naquela sentença, no exemplo apresentado no quadro abaixo: na sentença que possui um IP-SUB de número 41, o NP acusativo considerado é o de número 42 e o verbo - 45 (os números são atribuídos automaticamente pela ferramenta para a identificação do dado)

3) a sentença anotada

```
porque os erros facilmente se desculpam em favor de um morto; se bem que pouco vale
um livro, quando para merecer algum sufrágio, necessita que primeiro morra o seu Autor;
(AIRES-A_001,09.101)
*~/
/*
41 IP-SUB:  41 IP-SUB, 42 NP-ACC, 45 VB-P, 43 Q
*/
( (41 IP-SUB (42 NP-ACC (43 Q pouco))
             (45 VB-P vale)
             (47 NP-SBJ (48 D-UM um) (50 N livro))
             (52 , ,)
             (54 CP-ADV (55 WADV quando)
                        (57 IP-SUB (58 NP-SBJ *pro*)
                                   (60 PP (61 P para)
                                          (63 IP-INF (64 VB merecer)
                                                     (66 NP-ACC (67 Q algum) (69 N
sufrÃ¡gio))))
                                   (71 , ,)
                                   (73 VB-P necessita)
                                   (75 CP-THT (76 C que)
                                              (78 IP-SUB (79 ADVP (80 ADJ primeiro))
                                                         (82 VB-SP morra)
                                                         (84 NP-SBJ (85 D o) (87 PRO$ seu)
(89 NPR Autor)))))))
```

Quadro 2: exemplo de dado encontrado no arquivo de saída (output)
Fonte: Elaboração própria

No final do arquivo de saída, a busca trás automaticamente a quantificação *hits/tokens/total* – onde *hits* é o número de dados / *tokens* é o número de sentenças que contém dados / e *total* é o total de sentenças do(s) texto(s).

```
SUMMARY:
   source files,          hits/tokens/total
  ..\port\aires.psd          12/12/4498
whole search,          hits/tokens/total
                           12/12/4498
```

Quadro 3: exemplo da quantificação dada automaticamente no arquivo de saída (output)

A desvantagem de se utilizar o formato *TreeBank* e *Corpus Search* é que aquele formato ainda não pôde ser acoplado ao XML e portanto, a versão anotada sintaticamente é uma versão do texto separada da raiz XML, o que tem como consequência o fato de que algumas informações não poderão ser recuperadas, não podemos, por exemplo, realizar uma busca considerando a anotação sintática e poder escolher imprimir o dado ora no formato da transcrição paleográfica, ora no formato da edição modernizada, reduzindo assim as possibilidades de recuperação de informações e de visualização da busca que seriam permitidas se todas as informações estivessem encaixadas em um arquivo XML.

Aplicações WEB

No intuito de permitir que o corpus DOViC colabore com diversas pesquisas linguísticas, filológicas e históricas, está sendo desenvolvida[vi] uma aplicação baseada na WEB para disponibilizar os textos do corpus na Internet. Entendemos que essa disponibilização pode contribuir com a investigação da história do português brasileiro, uma vez que seu conteúdo será acessível ao público leitor em geral. Por se tratar de um corpus de documentos históricos, disponibilizá-lo juntamente com a fotografia dos documentos originais promoverá a divulgação e acesso ao patrimônio histórico, social e cultural de Vitória da Conquista, pois muitos pesquisadores não têm acesso material a esses documentos.

A aplicação fornecerá funcionalidades de buscas sintáticas e morfológicas automatizadas, provendo uma interface gráfica de elevada usabilidade para este propósito. Assim, o pesquisador pode concentrar seus esforços em seu objeto de estudo, dispensando a aprendizagem de linguagens/comandos específicos das ferramentas de busca.

O software permitirá a visualização do texto transcrito, das fotografias dos documentos originais, da ficha catalográfica, do léxico de edições e do texto editado em diferentes versões. O usuário poderá também realizar o download dos documentos tal como visualizados ou com as anotações em XML disponíveis para cada texto.

As tecnologias utilizadas no desenvolvimento desta aplicação baseiam-se na plataforma Java com o banco de dados PostgreSQL. A implementação das funcionalidades de busca utilizam tecnologias aplicáveis ao formato XML. As figuras 2 e 3 demonstram protótipos de algumas telas do sistema já produzidas como artefatos.

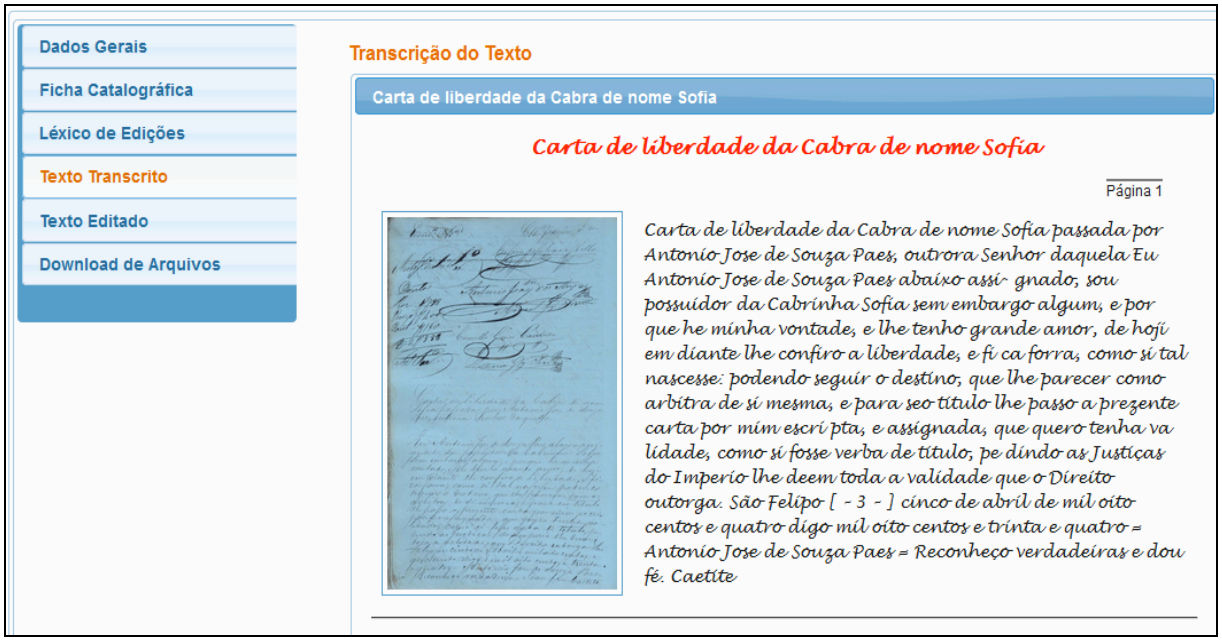

Figura 2: Tela do sistema Web para visualização do texto transcrito

Fonte: Elaboração própria (*print-screen* da tela do programa)

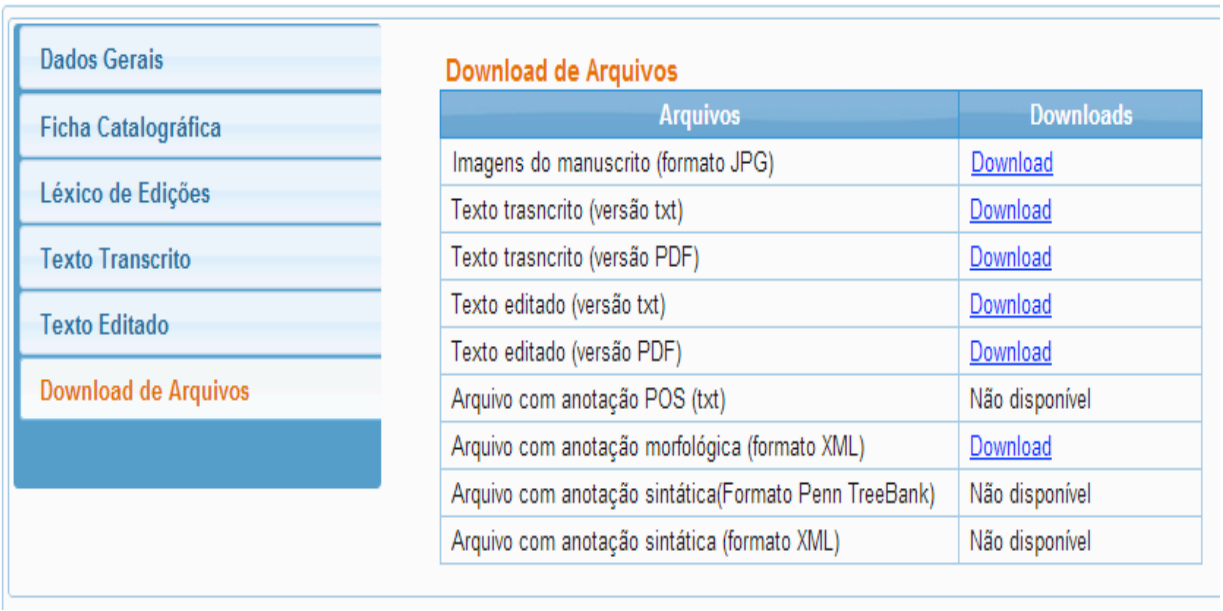

Figura 3: Tela do sistema Web para download de arquivos dos textos do corpus

Fonte: Elaboração própria (*print-screen* da tela do programa)

A importância de Sistemas de Informação

As pesquisas de modo geral caracterizam-se como projetos, que segundo o PMBOK (Project Management Body of Knowledge) são empreendimentos temporários com o intuito de desenvolver um produto ou serviço únicos. Devido à multiplicidade de recursos e atividades envolvidos no projeto de construção de um corpus, é importante que o trabalho seja gerenciado de maneira eficiente. Assim, além de técnicas e habilidades de gerenciamento, ferramentas de suporte ao planejamento, coordenação e controle são fundamentais para garantir a confiabilidade das informações, a produtividade, e otimização de tempo e esforços. No intuito de alcançar tais objetivos no projeto de construção do DOViC, Farias, Namiuti e Santos (2012) lançaram–se em um desafio: a construção de um Sistema de Informação[vii] que possibilite gerenciar o trabalho dos pesquisadores em um laboratório de pesquisa cientifica. Laudon e Laudon (2004) conceituam um sistema de informação como "um conjunto de componentes inter-relacionados que coleta (ou recupera), processa, armazena e distribui informações destinadas a apoiar a tomada de decisões, a coordenação e o controle de uma organização." O Sistema de Informação desenvolvido atende aos requisitos necessários para o projeto do corpus em questão, possuindo algumas características de software de Gestão de Projetos e Sistemas de Informação Gerenciais.

Visando o controle do trabalho para a construção do corpus, orientação das pesquisas e gerenciamento dos projetos, os professores coordenadores do Lapelinc/UESB, Jorge Viana Santos e Cristiane Namiuti-Temponi lançaram-se em um projeto para o desenvolvido o SGP (Sistema de Gerenciamento de Pesquisa)[viii] [um SI (sistema de informação), desenvolvido em JAVA, que comporta-se como um software de Gestão de Projetos e possui características de um SIG (Sistema de Informações Gerenciais). O sistema auxilia no gerenciamento dos projetos desenvolvidos no Lapelinc, permitindo a atribuição de tarefas e controle do andamento das atividades. Outras funcionalidades como gerenciamento de produções acadêmicas e apoio à decisão na compra de equipamentos também foram implementadas.

A interface do SGP ainda encontra-se em fase de ajustes, no entanto apresentamos neste artigo algumas ilustrações dessa interface em construção. As figuras 4 e 5 ilustram as janelas de atribuição e acompanhamento de tarefas do SGP.

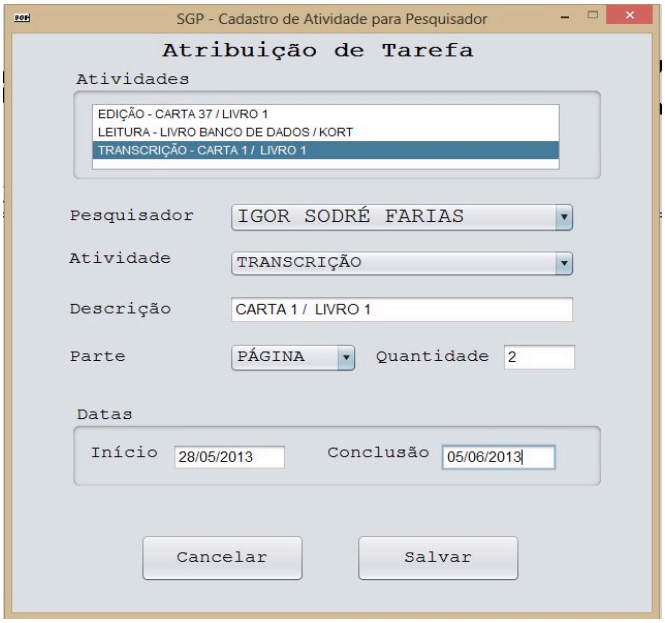

Figura 4 - Janela de atribuição de tarefas no SGP[ix]

Fonte: Elaboração própria (*print-screen* da tela do programa)

Figura 5 - Tela de registro de realização de tarefa[x]

Fonte: Elaboração própria (*print-screen* da tela do programa)

Como um sistema de Gerenciamento de Projetos, O SGP torna possível verificar o tempo gasto pelos bolsistas em realizar determinadas tarefas, o que auxilia no planejamento

dos projetos gerando estimativas de prazos cada vez mais exatas, provido pelo histórico de informações armazenadas.

Com as informações que o software pode armazenar e manipular, questões relevantes como quantidade de produções publicadas ao longo dos anos por pesquisadores que formam a equipe do laboratório poderão ser respondidas sem grande dificuldade, além de poder avaliar o crescimento de publicações nas várias edições de eventos científicos.

Outro tópico de bastante interesse para os coordenadores de projetos é o modulo do SGP que permite gerenciar o patrimônio do laboratório de pesquisa, uma vez que ele armazena todos os dados dos dispositivos inclusive em relação aos reparos que esses já sofreram. Assim, o SGP também se caracteriza como uma ferramenta de apoio à decisão, provendo informações que permitam uma melhor avaliação no momento da aquisição de novas ferramentas de trabalho evitando por exemplo, computadores que tiveram muitos problemas funcionais e privilegiando aqueles que tiveram um funcionamento de destaque positivo.

Assim, na exploração da fronteira representada pela união das esferas da Linguística de *Corpus* e da produção, organização e armazenamento de fontes e dados, o objeto "organização e gerenciamento da pesquisa" é também de suma importância. A necessidade de agilidade e automação na recuperação de informação pode ser suprida com sistemas de gerenciamento de informações, banco de dados e ferramentas de busca automática.

Considerações Finais

Pelo exposto, pode-se observar que, corroborando os pressupostos iniciais de Namiuti, Santos e Leite (2011), no âmbito dos projetos de pesquisa mencionados, as respostas desenvolvidas e em desenvolvimento tem se mostrado promissoras no sentido de potencializar a união da Linguística com as vertentes tecnológica e computacional, ou, a rigor, tecnológica-computacional. Promissoras, em primeiro lugar, porque, como vimos, usar a Fotografia controlada por um método, como postulou Santos (2013), tem produzido resultados relevantes para corpora, como o DOViC. Em segundo porque, no tocante tanto a possibilidades de processamento automático, quanto a ferramentas de anotação e busca (itens 3 e 4), projetos como os de Namiuti (2012), tem demonstrado, na prática, como pressupôs Paixão de Souza (2004, 2006), a importante vantagem para pesquisas com sintaxe, do texto

codificado e anotado eletronicamente, com o auxílio de linguagens de programação e de busca. E, em terceiro, porque experiências como a do Lapelinc, tem servido para demonstrar a importância, nem sempre lembrada, de que os resultados parciais do trabalho de pesquisa, durante e após o desenvolvimento como projeto, requer, de um lado, um gerenciamento eficaz, o que é, como vimos, o objetivo principal do SGP, destacado em Farias, Namiuti e Santos (2013); e, de outro, requer interfaces gráficas, a exemplo da proposta em Costa (2013).

Assim combinada com a tecnologia, a Linguística não só pode se beneficiar na eficácia de suas análises, como também abre um espaço ainda mais promissor para que o trabalho com o dado de língua não seja isolado da preocupação com a preservação daquilo que, sem dúvida, está contido nos textos: a memória e história tanto da língua quanto da sociedade.

Referências

CASTILHO, A. T. Para a História do Português Brasileiro. Projeto de pesquisa. USP, São Paulo, 1997.

CHOMSKY, **Principles and parameters in linguistic theory.** Cambridge (MA). MIT Press, 1979.

CORPUS SEARCH**. Corpus Search: Users Guide**. 2009. Disponível em: <http://*corpus*search.sourceforge.net/CS-manual/Contents.html>. Acesso em: 30 jan. 2013.

COSTA, A. S. **Uma aplicação Web para disponibilização e recuperação de informação do corpus digital DOViC**. Projeto de Mestrado em Lingüística. UESB, Vitória da Conquista, 2013. (Orientador: Cristiane Namiuti-Temponi; Co-orientador: Jorge Viana Santos).

FARIA, P.; KEPLER, F. N.; PAIXÃO DE SOUZA, M. C. An Integrated Tool for Annotating Historical Corpora , In: Fourth Linguistic Annotation Workshop, LAW IV, **48th Annual Meeting of the ACL**, 2010, Uppsala. Proceedings of the Fourth Linguistic Annotation Workshop, 2010. p. 217-221.

FARIAS, I. S. NAMIUTI, C. SANTOS, J. V. **Uso do banco de dados para a pesquisa de linguística: desenvolvimento tecnológico para a vertente computacional do corpus DOViC**. 2013. XVII Congresso de Iniciação Científica e Tecnológica.

GALVES, C. C. et al. **Padrões Rítmicos, fixação de parâmetros e mudança lingüística**. Projeto de pesquisa. UNICAMP, Campinas, 1997.

GALVES, C. C. O português no tempo e no espaço. Projeto de pesquisa. UNICAMP, Campinas, 2012 (FAPESP: 2012/06078-9).

LAUDON, K. C.; LAUDON, J. P. **Sistemas de Informação Gerenciais**. Administrando a

empresa digital. 5. ed. São Paulo: Prentice Hall, 2004.

MARCUS, M.; TAYLOR, ANN. **The Penn TreeBank Project**. Disponível em: <http://www.cis.upenn.edu/~treebank/> 2002. Acesso em 14 de jan. de 2013.

MATOS E SILVA, R. V. **PROHPOR** - Programa para a História da Língua Portuguesa. Projeto de pesquisa. UFBA, Salvador, 1992.

NAMIUTI, C. **Novos meios para antigas fontes: Sintaxe Diacrônica em *corpus* eletrônico do português**. Projeto de Pesquisa. UESB, Vitória da Conquista, 2010. (Convênio FAPESB/UESB 006/2012)

NAMIUTI, C.; SANTOS, J. V.; LEITE, C. M. B. Propostas e Desafios dos Novos Meios das Antigas Fontes: A Preservação da Memória pela Linguística de *Corpus*. In: X Colóquio Nacional e II Colóquio Internacional do Museu Pedagógico UESB, 2011, Vitória da Conquista. **Anais do X Colóquio Nacional e II Colóquio Internacional do Museu Pedagógico** UESB. Vitória da Conquista: UESB, 2011. v. 1. p. 1-11.

PAIXÃO DE SOUSA, M. C. (2004). **Memórias do Texto**. Projeto de Pesquisa. FAPESP-UNICAMP, Campinas, 2004.

PAIXÃO DE SOUSA, M. C. Digital Text: Conceptual and methodological frontiers. In: ROMERO, D.; SANZ, A. (Org.). **Literatures in the Digital Era: Theory and Praxis**. Cambridge: Cambridge Scholarly, 2007.

PAIXÃO DE SOUSA, M. C. Memórias do Texto. **Revista Texto Digital**, n. 2., 2006. Disponível em: <http://www.textodigital.ufsc.br/num02/paixao.htm>. Acesso em: 01 out. 2012.

PAIXÃO DE SOUZA, M. C.; KEPLER, F.N.; FARIA, P. E-Dictor: novas perspectivas na codificação e edição de corpora de textos. In: SHEPHERD, Tania; SARDINHA, Tony Berber; PINTO, Marcia Veirano (Org.). **Caminhos da Linguística de *corpus***. Campinas: Mercado de Letras, 2010.

SANTORINI, B. **Annotation manual for the Penn Historical *Corpora* and the PCEEC**. Disponível em: <http://www.ling.upenn.edu/hist-*corpora*/annotation/index.html>. 2010. Acesso em 08 de jan. de 2013.

SANTOS, J. V. **Apresentação de meios para o transporte: digitalização de documentos manuscritos e impressos**. Conferência ministrada na I Oficina de Lingüística de *Corpus* da Bahia (UEFS, UESB, UFBA). Feira de Santana: UEFS, 2010a.

SANTOS, J. V. **Memória da escravidão baiana: análise semântica comparativa de sentidos de liberdade em cartas de alforria oitocentistas de Vitória da Conquista e Rio de Contas.** Projeto de Pesquisa. UESB, Vitória da Conquista, 2012. (Convênio FAPESB/UESB 006/2012).

SANTOS, J. V. O ponto de vista semiótico na fotografia rodchenkiana. In: XV Congresso Brasileiro de Ciências da Comunicação, 2002, Salvador. **Anais do XV Congresso Brasileiro de Ciências da Comunicação**, 2002.

SANTOS, J. V. **Técnicas de transporte do texto manuscrito para o meio digital**. Conferência ministrada na I Oficina de Lingüística de *Corpus* da Bahia (UEFS, UESB, UFBA). Feira de Santana: UEFS, 2010b.

SANTOS, J. V. **Um método de Fotografia técnica documental para formação de corpora digitais de documentos históricos manuscritos**. 2013. (No prelo.)

SANTOS, J. V.; NAMIUTI, C. **Corpora Digitais Para a História do Português Brasileiro - região Sudoeste da Bahia: Aliança PHPB - Tycho Brahe.** Projeto de pesquisa. UESB, Vitória da Conquista, 2010. (FAPESB: 6171/2010)

SANTOS, J. V.; NAMIUTI, C. **Memória conquistense**: recuperação de documentos oitocentistas na implementação de um *corpus* digital. Projeto de pesquisa. UESB, Vitória da Conquista, 2009.

SANTOS, J.V. **Memória Conquistense: recuperação de documentos oitocentistas na implementação de um *corpus* digital**. Projeto de Pesquisa. UESB, Vitória da Conquista, 2009.

UNICAMP. ***Corpus* Histórico Anotado do Português Tycho Brahe**. 1998. Disponível em: <www.tycho.iel.unicamp.br/~tycho/corpus>. Acesso em: 03 jan. 2013.

W3C. **XML Path Language (**XPath**)**. 1999. Disponível em: http://www.w3.org/XML Acesso em: 30 jan. 2012.

NOTAS

[i] Gráficos originalmente em Inglês (cf. SANTOS, 2013).
[ii] Gráficos originalmente em Inglês (cf. SANTOS, 2013).
[iii] O Corpus Histórico do Português Tycho Brahe é um corpus eletrônico anotado, composto de textos em português escritos por autores nascidos entre 1380 e 1845. Disponível em: http://www.tycho.iel.unicamp.br/~tycho/corpus/index.html
[iv] Relação estrutural definida pela Teoria da Gramática Gerativa, Noam Chomsky (1982).
[v] Etiqueta prevista no sistema de anotação sintática do corpus Tycho Brahe para designar a categoria gramatical equivalente a "oração" (cf. http://www.tycho.iel.unicamp.br/corpus/manual/syn-frm.html)
[vi] No âmbito do projeto de mestrado de Aline Silva Costa: Uma aplicação Web para disponibilização e recuperação de informação do corpus digital DOViC. 2013. (Mestrado em LINGUÍSTICA) - Universidade Estadual do Sudoeste da Bahia. Cristiane Namiuti-Temponi (Orientador). Jorge Viana Santos (Co-orientador).
[vii] Um Sistema de Informação Gerencial gera relatórios que ajudam administradores a melhorar a elaboração de planos e tomada de decisões e a obter um maior controle sobre as operações de sua organização. (REYNOLDS; STAIR, 2006).
[viii] O SGP foi idealizado na ano de 2012 pelos professores Jorge Viana Santos (UESB) e Cristiane Namiuti-Temponi (UESB) e programado pelo bolsista de Iniciação científica Igor Sodré Farias sob a orientação dos professores.
[ix] Nessa janela é possível atribuir uma tarefa a um pesquisador e registrar o tempo de realização do trabalho.
[x] Essa janela permite o pesquisador cadastrar o tempo gasto para executar a tarefa.

Sobre os autores

Cristiane Namiuti Temponi é Doutora em Lingüística pela Universidade Esdadual de Campinas (UNICAMP). Tem experiência na área de Lingüística, com ênfase em Lingüística Histórica e metodologias automáticas de busca de dados em textos escritos, atuando principalmente, nos seguintes temas: interpolação, clítico, mudança linguística, história do português e linguistica de corpus. Possui Bacharelado em Linguística pela UICAMP (2001), Doutorado (2008) e Pós-Doutorado (2010), em Lingüística, pela mesma instituição. Atualmente é professora da Universidade Estadual do Sudoeste da Bahia (UESB) e professora do quadro permanente do Programa de Pós-Graduação em Linguística (PPGLin-UESB).

Jorge Viana Santos é Doutor em Lingüística pela Universidade Esdadual de Campinas (UNICAMP), Mestre em Comunicação e Semiótica pela Pontifícia Universidade Católica de São Paulo. Atualmente é professor Adjunto da Universidade Estadual do Sudoeste da Bahia e professor do quadro permanente do Programa de Pós-Graduação em Linguística (PPGLin-UESB) e docente colaborador do Programa de Pós-Graduação em Memória: Linguagem e Sociedade (PPGMLS/UESB). Tem experiência na área de Lingüística e Semiótica, atuando nos seguintes temas: sentido, argumentação, lugares de enunciação, processos de designação, reescritura, subjetivação, textos, Linguística de Corpus, fotografia, imagem e memória.

Aline Silva Costa é estudante de pós-graduação em Lingüística na Universidade Estadual do Sudoeste da Bahia (UESB) sob a orientação e co-orientação dos Professores: Cristiane Namiuti-Temponi e Jorge Viana Santos. Possui graduação em Ciência da Computação pela UESB (2004) e especialização em Administração de Sistemas de Informação pela Universidade Federal de Lavras (2005). Atualmente é professora efetiva do Instituto Federal de Educação, Ciência e Tecnologia da Bahia. Tem experiência na área de Ciência da Computação, com ênfase em Metodologia e Técnicas da Computação.

Igor Sodré Farias é estudante do Curso de Graduação em Ciência da Computação na Universidade Estadual do Sudoeste da Bahia. Desenvolve projeto de Iniciação Científica (Bolsista CNPq) na Área de Lingüística de Corpus, contribuindo com a criação de sistemas de gerenciamento de informação para a construção do corpus de Documentos Oitocentistas de Vitória da Conquista (DOViC), trabalho orientado e co-orientado pelos Professores Cristiane Namiuti-Temponi e Jorge Viana-Santos.