ABHD
Revista Brasileira de Humanidades Digitais RBHD

UMA PROPOSTA METODOLÓGICA PARA A CONSTRUÇÃO DE CORPORA ATRAVÉS DE ESTRUTURAS DE TRABALHO: O LAPELINC FRAMEWORK

A methodological proposal for the construction of corpora through work structures: LAPELINC Framework

Bruno Silvério Costa,Jorge Viana Santos,Cristiane Namiuti

Resumo. *No âmbito das Humanidades Digitais, a construção de corpora tem sido objeto da Linguística de Corpus, que baseada em demandas específicas e oriundas de necessidades particulares de cada pesquisa, tem gerado resultados com interface limitada. Assim, identifica-se uma lacuna existente nas iniciativas de produção de corpora, resultado da forma "exploratória" de condução dessas mesmas iniciativas. Buscando uma solução que possibilite obter os benefícios da construção padronizada de corpora, a fidedignidade aos documentos originais, a consistência filológica e a interoperabilidade entre iniciativas de pesquisa, propõe-se neste trabalho uma metodologia para a criação de corpora linguísticos: o LAPELINC Framework.*

Abstract. *In order for us to prepare the work for online publishing, the author must be careful of the submission guidelines. Papers in Portuguese should present an abstract in Portuguese and abstract in English and papers in Spanish, an abstract in Spanish and an abstract in English. For papers in English, only an abstract in English should be provided. Abstracts should not exceed 10 lines each and have to be on the first page.*

Introdução

As Humanidades Digitais constituem um modelo de investigação que infere a utilização dos meios de processamento digital da informação como instrumentos de transformação metodológica, empregados na produção de conhecimento nas humanidades (PORTELA, 2013).

Sob essa óptica, surge a possibilidade de se observar a língua a partir da lente de ferramentas computacionais, utilizando grandes repositórios de dados como objeto de estudo, associados a metadados e programas que os manipulam, denominados corpora eletrônicos (FREITAS, 2017).

No âmbito das Humanidades Digitais, a construção de corpora tem sido objeto da Linguística de Corpus, uma subárea da Linguística Computacional. Envolvendo vários processos complexos, a construção de corpora compreende as etapas de coleta, preparação, segmentação e anotação de textos (EVANS, 2008). Essa referida construção é normalmente baseada em demandas específicas e oriundas de necessidades particulares de cada pesquisa, gerando resultados com interface limitada com outras iniciativas, além da baixa possibilidade de reuso das soluções desenvolvidas (ALVES; MORINAKA, 2014).

Compilar um corpus, além de manter e disponibilizar acesso aos dados e informações nele armazenados, é uma tarefa criteriosa e fundamentada em princípios e teorias que demandam rigor em sua execução (MENDES, 2016). Conforme Aluisio e Almeida (2006), evocando os conceitos apresentados por McEnery e Wilson (1996), a moderna noção de corpus apresenta quatro características fundamentais: amostragem e representatividade, tamanho finito, formato eletrônico e referência padrão (ou reuso do corpus). Dessa forma, tanto o formato eletrônico quanto a possibilidade de reuso tornam um corpus dependente das tecnologias da informação para sua implementação e manutenção.

Construir corpora de documentos históricos necessita da garantia de fidedignidade. Esse princípio basilar demanda que componentes das dimensões computacional e linguística estejam corretamente integrados, possibilitando ao meio digital a adequada representação do objeto real. Na interface entre as dimensões computacional e filológica, encontra-se o desafio de desenvolver o tratamento de textos que permita o processamento automático, sem contudo perder a consistência filológica do material (PAIXÃO DE SOUSA, 2006).

Como parte integrante ao processo de implementação de corpora de textos antigos, 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, tem desenvolvido e aplicado o método LAPELINC como um fluxo de trabalho (*workflow*) que permite a transposição do documento físico para o meio digital, mantendo a fidedignidade desses documentos aos seus originais físicos. No método LAPELINC, após as etapas de transposição e transcrição, o texto é portado para o formato digital de texto simples (TXT), requisito necessário aos processos que se

seguem. Esse texto simples é então editado e anotado linguisticamente (anotações morfossintática e sintática), sendo portado para o formato XML que representa as diversas camadas de anotação em um único arquivo. Esse conjunto de informações de camadas distintas associadas ao texto permitem ao pesquisador selecionar cada uma das diferentes visões disponíveis. (SANTOS; NAMIUTI, 2016; NAMIUTI; SANTOS, 2016).

Padronizar técnicas e procedimentos permite estabelecer uma forma *standard* para resolver problemas de uma mesma tipologia, ou com características similares, selecionando práticas que diminuem a complexidade na realização das atividades, além de reduzir as chances de erro e o índice de retrabalho (MAGALHÃES; PINHEIRO, 2007).

Assim, identifica-se uma lacuna existente nas iniciativas de produção de corpora, resultado da forma "exploratória" de condução dessas mesmas iniciativas. Embora exista particularidades pertinentes a cada pesquisa, é possível identificar uma série de etapas comuns na construção de corpora, estabelecendo um percurso essencial compartilhado. Respeitando a necessidade de adaptação às peculiaridades de cada pesquisa e ao mesmo tempo, buscando uma solução que possibilite obter os benefícios da construção padronizada de corpora, a fidedignidade aos documentos originais, a consistência filológica, com possibilidade direta de redução do custo de implementação do corpus e o reuso de corpora produzidos, propõe-se neste trabalho, como recorte da pesquisa da tese de doutoramento em andamento de Costa (2019) no âmbito do PPGLIN/UESB, uma metodologia para a criação de corpora linguísticos: o LAPELINC Framework.

Metodologia

O desenvolvimento da solução apresentada neste trabalho buscou conservar nas diversas etapas da metodologia desenvolvida, a consistência filológica e a fidelidade da representação digital em relação ao original, conforme apresentados respectivamente por Paixão de Sousa (2006) e Namiuti e Santos (2016). As funcionalidades requeridas ao projeto de corpora basearam-se nos princípios metodológicos de Aluíso e Almeida (2006).

Para representar a complexidade do objeto livro para corpora baseados em documentos físicos, foram observados os princípios estabelecidos em Santos e Namiuti (2017), seguindo o *workflow* do método LAPELINC para a compilação de corpora (NAMIUTI; SANTOS, 2016).

A definição do processo de transcrição paleográfica baseou-se na obra de Berwanger e Leal (2015). Os parâmetros para o georreferenciamento de textos e buscas em estudos de geolingüística seguiram os fundamentos apresentados em Cardoso (2010).

Foram utilizadas as diretrizes apresentadas por Aho et al. (2007) para a proposta computacional dos módulos de *tagging* e *parsing* sintático, seguindo um percurso adequado à automatização requerida (AUROUX, 1998, p. 466), espelhando-se e na metodologia de anotação utilizada para a representação de metadados do Corpus Tycho Brahe (GALVES, 2019; PAIXÃO DE SOUSA, 2014). Para representar a informação semântica durante o processo de anotação, utilizou-se das técnicas e princípios apresentados na teoria da Semântica do Acontecimento (GUIMARÃES, 2017).

As funcionalidades de reuso e a disponibilidade dos corpora implementados para acesso por outros pesquisadores, basearam-se nas características apresentadas por McEnery e Wilson (2001) e Aluísio e Almeida (2006).

A proposta do LAPELINC Framework apresenta uma estrutura genérica, compatível com a utilização de representações, anotações e ferramentas diversos, contendo tanto partes fixas (*frozenspots*), como flexíveis (*hotspots*) (CROSSAN; LANE; WHITE, 1999), além de obedecer a padrões de construção para frameworks de software apontados por Fayad, Schmidt e Johnson (1999).

Pressupostos teóricos

Definida por Teubert (1996) como “a face moderna da linguística empírica”, a Linguística de Corpus provê recursos poderosos para as investigações teóricas sobre a linguagem. Para McEnery e Wilson (MCENERY; WILSON, 2001), o próprio conceito de corpus tem mudado com o passar do tempo, assumindo nas últimas décadas um significado moldado pela Linguística de Corpus, definido por quatro características fundamentais:

a) amostragem e representatividade – o tamanho do corpus deve ser suficientemente grande para representar o fenômeno estudo em uma época e local determinados;
b) tamanho finito – apesar de ser suficientemente grande para representar um recorte dos falantes de uma determinada época e lugar, o corpus deve poder ser delimitado, ou seja, as fronteiras que dizem o que pertencem ou não a ele devem ser nítidas;

c) formato eletrônico – os textos contidos em um corpus devem assumir o formato eletrônico, ou seja, o suporte para o texto deve utilizar uma ou mais das tecnologias digitais disponíceis. Esse formato permite ao corpus i) ser pesquisado de forma rápida e ii) ter seu valor aumentado a partir da agregação de novas informações de forma facilitada;

d) referência padrão – um corpus constitui-se uma referência para os fatos de língua que ele representa, pressupondo a sua disponibilização a outros pesquisadores, possibilitando o chamado de reuso do corpus.

Para Aluísio e Almeida (2006), a possibilidade de reuso é "digna de nota", uma vez marca nitidamente a diferença entre o conceito de corpus para a Linguística e para a Linguística de Corpus. Ainda segundo Aluísio e Almeida (2006), disponibilizar um corpus para uso futuro torna-se uma característica fundamental do corpus eletrônico. Isso torna o esforço empreendido para a construção de um corpus relacionado a uma pesquisa específica contribuinte ao trabalho realizado por terceiros sobre esse mesmo corpus, uma vez que serve como base de referência para a língua por ele representado. Os textos a serem considerados pela Linguística de Corpus devem assumir dimensões conhecidas, além de serem organizados obedecendo a um projeto que assuma uma perspectiva linguística adequada.

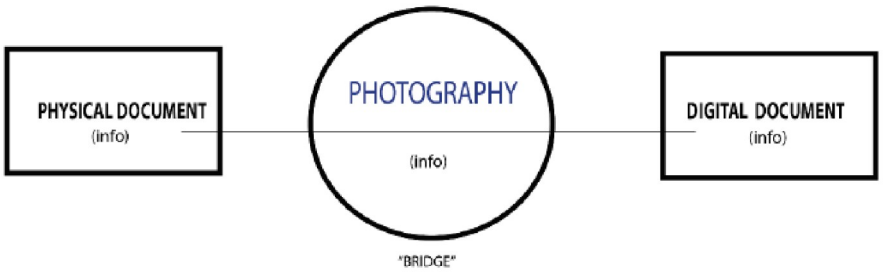
A análise linguística efetivada através de meios automáticos, demanda que os textos a serem examinados sejam previamente preparados, apresentando formatação e codificação próprias do meio digital, impondo limites à representação da grafia e de outras características peculiares ao documento original. Considerando o estudo de documentos antigos, cujas características distinguem-se grandemente das formas hodiernas de suporte, evidencia-se a necessidade de garantir a fidedignidade ao documento original físico, conforme estabelecem parâmetros filológicos relevantes a serem respeitados na seleção e classificação dos textos incluídos no corpus. Conforme a temática de cada pesquisa, são necessárias pequenas alterações aos textos originais, as edições, podendo trazer impactos consideráveis à análise linguística. O equilíbrio entre as dimensões computacional e filológica desta problemática, isto é, o tratamento automático de textos de forma filologicamente consistente, resulta em um desafio a ser superado (PAIXÃO DE SOUSA, 2006).

O suporte material do livro físico limita seu acesso a um tempo e a um espaço determinados (SANTOS; NAMIUTI, 2016). Cercado de complexidades, o documento físico impõe limitações à sua manipulação, restringindo o acesso ao mesmo a um pequeno conjunto de pesquisadores. No decurso da história, tecnologias e meios adequados ao suporte material foram criados,

garantindo a preservação e o resgate de informações e documentos. O suporte material para o texto estabelece uma forma de se fazer humanidades, sendo o desenvolvimento tecnológico humano gerador de tipos distintos de suportes para as fontes documentais. O meio eletrônico, representado pelas tecnologias digitais, trazem consigo uma nova forma de suporte, além de novas e promissoras possibilidades, definindo também uma nova forma de se fazer humanidades denominada de Humanidades Digitais. Essas complexidades são impactadas pela utilização do meio digital, uma vez que o acesso e a manipulação remotos permitem a disponibilização de documentos sem a possibilidade de danos ao original físico, além de ampliar o possível número de leitores (SANTOS; NAMIUTI, 2017). Nesse sentido, foi desenvolvido e aplicado uma metodologia de construção de corpora digitais anotados e cientificamente controlados denominada Método LAPELINC (NAMIUTI; SANTOS, 2016).

No Método LAPELINC, o Documento Físico (DF) é utilizado para construir um Documento Digital Imagem (DDI), conforme apresentado na Figura 1, e que funcionará como fonte original no meio digital para a construção de corpora eletrônicos anotados, compostos por arquivos do tipo Documento Digital Texto (DDT).

Figura 1 - Fotografia praticada com método científico de reprodução digital: a ponte entre Documento Físico e Documento digital



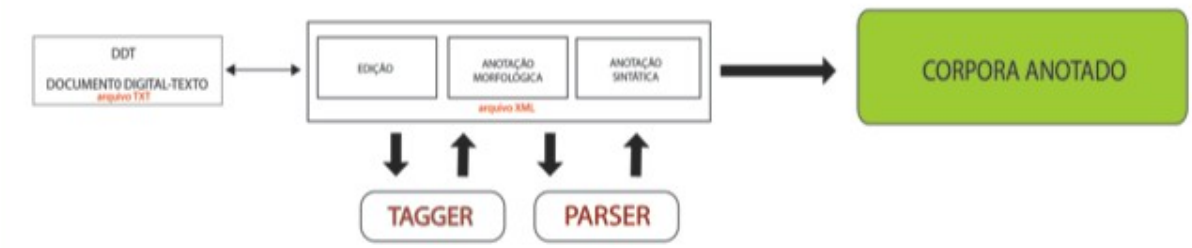
Fonte: Namiuti; Santos *et al.*, (2013)

Na transposição, o método LAPELINC apresenta um aparato que garante a recuperação de informações variadas, denominado Aparato de Metadados Estruturados (AME). O AME permite recuperar, no digital, a complexidade do Documento Histórico Físico, "cuja realidade é tridimensional, irregular e subjetivamente construída" (SANTOS; BRITO, 2017). O AME é formado por cinco componentes básicos: i) Catálogo Visual (CV); ii) Dossiê de Observações Pertinentes (DOP); iii) Fotografia Cientificamente Controlada (FCC); iv) Análise Topográfica (AT) e v) Análise Descritiva (AD). Atualmente, o AME engloba cinco componentes fundamentais: i)

Catálogo Visual (CV); ii) Dossiê de Observações Pertinentes (DOP); iii) Fotografia Cientificamente Controlada (FCC); iv) Análise Topográfica (AT); v) Análise Descritiva (AD) (SANTOS; NAMIUTI, 2017; COSTA; NAMIUTI; SANTOS, 2019; NAMIUTI; SANTOS, 2016)

Para permitir a utilização linguística adequada ao propósito específico de cada pesquisa, os documentos obtidos e tratados pelo Método LAPELINC, agora já em formato de texto puro após uma etapa de Transcrição Paleográfica, precisam ser processados em programas especializados para agregar metainformações aos mesmos. Após o processamento dos referidos textos em programas que possibilitem a execução das etapas de edição, anotação morfológica e anotação sintática, obtém-se finalmente o corpus eletrônico anotado, conforme apresentado na Figura 2 (NAMIUTI; SANTOS, 2016).

Figura 2 - Representação gráfica da etapa de Compilação de corpora do fluxo de trabalho do Método LAPELINC.



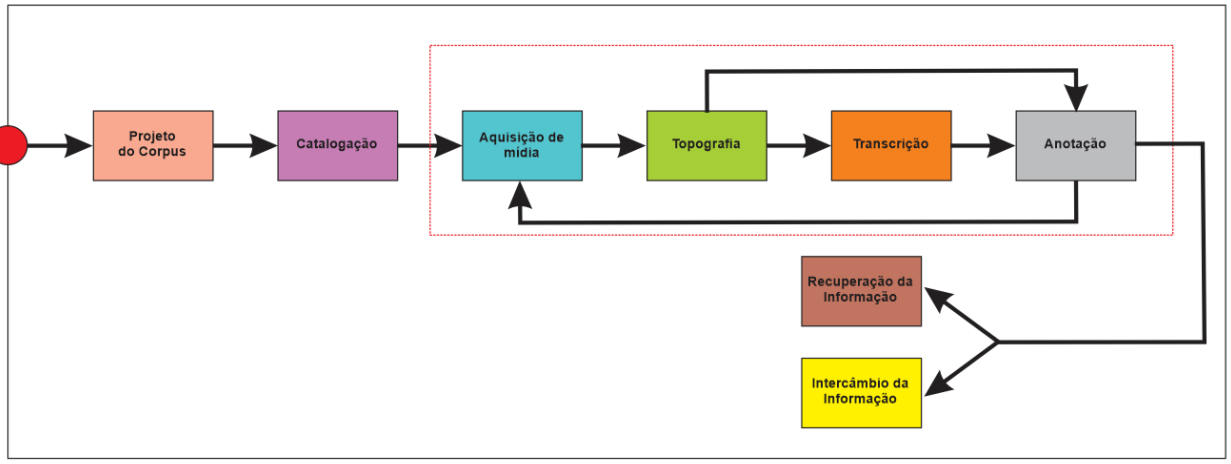
Fonte: Adaptado de Namiuti e Santos (2016)

Lugar comum dentro da área de Computação, a padronização de técnicas e procedimentos permite a geração de uma linguagem universal sobre um determinado tema. O estabelecimento de melhores práticas diminui a complexidade na realização das atividades, além de reduzir as chances de erro e o índice de retrabalho (MAGALHÃES; PINHEIRO, 2007). Um *Framework*, termo originário na língua inglesa e que pode ser traduzido como "estrutura", é amplamente empregado como conceito chave relacionado à criação de diretrizes de trabalho em diversas áreas de conhecimento. Para a Ciência da Computação, um Framework é uma estrutura de trabalho que tem como principal objetivo resolver problemas recorrentes com uma abordagem genérica, possibilitando a concentração de esforços pautados nas melhores práticas conhecidas. (ISACA, 2013). Um framework estabelece um conjunto de ferramentas e *workflows*, possibilitando o desenvolvimento de padrão de execução de atividades, associado ao reuso de soluções previamente desenvolvidas (FAYAD; SCHMIDT; JOHNSON, 1999).

O LAPELINC Framework

A partir das referências apresentadas e da metodologia informada, e baseando-se no LAPELINC *Workflow* (NAMIUTI; SANTOS, 2016), foi estabelecido um conjunto de etapas comuns às pesquisas linguísticas que desenvolvem corpora, conforme apresentado na Figura 3.

Figura 3 - Etapas do LAPELINC Framework



Fonte: Elaboração dos autores

A etapa inicial é a de Projeto do Corpus. As demais etapas do Framework a serem cumpridas dependem necessariamente das demandas particulares de cada pesquisa. O framework admite ainda o trabalho de forma linear, isto é, cada etapa é executada uma após a outra e uma única vez, ou de forma iterativa e incremental, permitindo que o conjunto de etapas seja executado o número de vezes que for necessário.

Essas etapas correspondem desde ao processo de projeto do corpus, até as fases de tratamento por anotação e edição, além da recuperação e intercâmbio das informações registradas. Internamente, cada etapa é formada por tarefas fixas (*frozenspots*) e tarefas flexíveis (*hotspots*). As tarefas fixas tratam-se de diretrizes indispensáveis estabelecidas pelo *framework,* apresentando padrões de solução para problemas conhecidos, sendo previamente desenvolvidos e disponibilizados para o reuso. Já as tarefas flexíveis disponibilizam a customização necessária do *framework* para adequar-se às particularidades de cada pesquisa.

O LAPELINC Framework estabelece as seguintes etapas para a construção de corpora, a serem implementadas através de módulos do software:

i. **Projeto do corpus** – possibilita a definição dos parâmetros estabelecidos para a composição do corpus, implementando as seguintes tarefas:
    a. Definição de escopo – possibilita o registro de parâmetros referentes ao tamanho, gêneros textuais e abrangência histórica do corpus;
    b. Definição de modalidade ou tipo do corpus – especifica o tipo de documento original utilizado como fonte para a composição do corpus,

como: texto de documentos físicos, fac-símiles, áudio, vídeo ou texto de origem digital (arquivos ou web).

ii. **Catalogação** – permite a especificação do catálogo dos documentos a serem incluídos no corpus, além de possibilitar o registro de informações linguísticas e codicológicas relacionadas aos documentos. Estão incluídas neste módulo as seguintes funcionalidades: criação do catálogo, elaboração do Dossiê de Observações Pertinentes (DOP) especificado pelo Método LAPELINC (NAMIUTI; SANTOS, 2016), descrição codicológica e registro de informações sobre geolocalização (locais onde foi escrito e onde se encontra o original atualmente).

iii. **Aquisição de mídia** - que permite a obtenção de arquivos e dados para a construção de textos a serem incluídos no corpus. Especifica 5 (cinco) conjuntos de rotinas diferentes para atender aos tipos distintos de corpus. São elas:

   a. Rotina para Transposição (DF-DDI) – possibilita a execução da transposição de documentos físicos através dos parâmetros especificados pelo Método LAPELINC (NAMIUTI; SANTOS, 2016);
   b. Rotina para Aquisição de Imagens Digitais (fac-símiles) – permite a importação de imagens de fac-símiles existentes , além de possibilitar a edição e a sintetização de metainformações sobre a imagem, conforme parâmetros do Método LAPELINC (NAMIUTI; SANTOS, 2016);
   c. Rotina para Aquisição de Áudio - disponibiliza operações de importação e edição para a formação de um corpus de áudio;
   d. Rotina para Aquisição de Vídeo - disponibiliza operações de importação e edição para a formação de um corpus de vídeo;
   e. Rotina de Aquisição de Texto em Formato Digital – permite a importação direta de arquivos de texto em formatos determinados ou a exploração de sítios na Web em níveis especificados de recursão para a composição de fonte textual;

iv. **Topografia** – possibilita descrever a estrutura topográfica de um documento, seguindo diretrizes estabelecidas pelo Método LAPELINC (SANTOS; NAMIUTI, 2016). Estão incluídas as seguintes rotinas:

   a. Rotina para Cadastro de Documentos Macro (DMAs);

b. Rotina para Cadastro de Documentos Micro (DMIs);
c. Rotina para Compilação de Sumário.

v. **Transcrição** – disponibiliza recursos e ferramentas para a transcrição em texto das imagens, vídeos ou áudios. Apresenta três conjuntos de rotinas específicas para esta fase:
   a. Rotinas de Transcrição de Imagem – possibilita as atividades de leitura e decifração de documentos, transcrição paleográfica, alinhamento texto-imagem e consulta a dicionários internos de letras, grafia, abreviaturas e símbolos;
   b. Rotinas de Transcrição de Áudio – apresenta funcionalidades para a audição e transcrição, além de alinhamento texto-áudio;
   c. Rotinas de Transcrição de Vídeo – apresenta funcionalidades para visualização, audição e transcrição, além de alinhamento texto-vídeo.

vi. **Anotação** – possibilita o registro de metainformações de edição, morfossintaxe e semântica relacionados ao texto do corpus. Inclui os seguintes módulos internos:
   a. Módulo de POS Tagging – realiza a marcação dos textos do corpus através de *tags* POS (*Parts of Speech*), utilizando para isso programas internos ou externos vinculados ao *Framework*;
   b. Módulo de Parsing Sintático – realiza a anotação dos textos do corpus através de operações do *parsing* sintático internos ou externos vinculados ao *Framework*;
   c. Módulo de Edição – possibilita a correção manual das anotações automáticas realizadas pelos dois módulos anteriores, permitindo o registro de anotações de edição linguística, como junção, fragmentação e expansão;
   d. Módulo de Anotação Semântica – permite a seleção de uma das teorias semânticas disponíveis no *framework*, além da anotação da análise semântica do texto selecionada.

vii. **Recuperação da Informação** – permite a obtenção de informações linguísticas através de operações de busca e visualização nos textos do corpus. Inclui os seguintes módulos internos:

a. Módulo de buscas – possibilita realizar procedimentos de busca dos tipos textual, morfológica, sintática, semântica e georreferenciada sobre o corpus;
b. Módulo de visualização – disponibiliza opções de visualização específicas para o tipo de informação exibida. São operações pertinentes ao módulo: visualização das marcações (tags) e edições (original, modernização), composição e visualização de resultado de busca (através de construções gráficas da busca e resultados, além de possibilidades textuais), visualização do catálogo (sumário, miniaturas de fac-símiles, estruturas arbóreas), visualização da anotação semântica e visualização de obra em formato 3D (com possibilidade de inclusão de metainformações em Realidade Aumentada ou Virtual).

viii. **Intercâmbio da Informação** – possibilita a importação/exportação de textos do corpus, permitindo o compartilhamento dos corpora com outras iniciativas de pesquisa, acesso a recursos de software por outros corpora externos, além de operacionalizar funções de importação de corpora externos. Utilizando-se de tecnologias de serviço web, disponibiliza as operações relacionadas, proporcionando o compartilhamento de recursos com a comunidade científica. Possui dois módulos internos:

a. Módulo de interoperabilidade – possibilita a disponibilização de recursos dos corpora e rotinas implementadas pelo LAPELINC Framework para usuários de iniciativas externas;
b. Módulo de Importação/Exportação – possibilita a geração de cópias dos corpora cadastrados, possibilitando a criação de cópias de segurança e o compartilhamento com outras iniciativas de pesquisa. A partir do fornecimento de arquivos exportados, permite a importação de corpora para restauração e/ou utilização de corpora criados por terceiros.

O conjunto de módulos principais e atividades associados às etapas que pertencem ao escopo do LAPELINC Framework são apresentadas de forma consolidada no Quadro 1:

Quadro 1 - Atividades englobadas pelo LAPELINC Framework

Etapas	Módulo	Atividades

Projeto do Corpus	Projeto	Definição de escopo do corpus: tamanho, gêneros textuais, período histórico	Definição de Modalidade: texto, audio, vídeo, texto de origem digital (web)	Catalogação	Elaboração do DOP (Dossiê de Observações Pertinentes)	Descrição codicológica	Registro de geolocalização atual e da elaboração da obra
Aquisição de mídia (Imagem / Áudio / Vídeo / Texto digital)	**Transposição (DF-->DDI)**	Análise Descritiva	Captura	Edição (Recorte; Formatos: RAW, JPG e Tumbnail)	Co-indexação de imagens		
	Aquisição de Imagens Digitais	Importação de arquivo de imagem	Edição (Recorte; Formatos: RAW, JPG e Tumbnail)	Sintetização de imagem com carimbo de metadados			
	Aquisição de Áudio	Importação de arquivo de áudio					
	Aquisição de Vídeo	Importação de arquivo de vídeo					
	Aquisição de Texto	Importação de arquivo de texto (pdf, txt, html)					
Topografia	**Análise Topografica Descritiva**	Cadastro de documentos macro	Cadastro de documentos micro	Compilação de sumário			
Transcrição	**Transcrição de Imagens**	Leitura e Decifração	Transcrição Paleográfica	Alinhamento texto e imagem	Dicionário de Letras, grafias, abreviaturas e símbolos		
	Transcrição de Áudio	Audição e transcrição	Alinhamento texto e áudio				
	Transcrição de Vídeo	Visualização, audição e transcrição	Alinhamento vídeo e texto				

Anotação	**POS Tagging**	Compatibilidade do modelo de anotação e o tagger	Instalação do Tagger	Tagging			
	Parsing Sintático	Compatibilização entre o modelo de anotação e o parser	Vinculação ao parser	Treinamento do parser	Parsing		
	Edição	Correção de anotações meta-textuais	Modernização, junção, fragmentação, expansão de abreviaturas	Tradução			
	Anotação Semântica	Seleção de teoria semântica	Anotação semântica				
Recuperação da Informação	**Buscas**	Busca textual	Busca Morfológica	Busca Sintática	Busca Semântica	Busca Georreferenciada	
	Visualização	Visualização da Edição (estilo planilha, eDictor Clássico ou Popup de Tokens	Visualização da expressão de busca (árvore, expressão regular, modelo visual)	Visualização dos documentos cadastrados (lista, miniaturas, árvores)	Definição e recuperação da diagramação	Visualização 3D	
Intercâmbio da Informação	**Interoperabilidade**	Instalação de webservices	Configuração de webservices				
	Importação / Exportação	Importação de documentos	Exportação de documentos				

Fonte: Elaboração dos autores

Considerações Finais

O LAPELINC Framework fornece um arcabouço metodológico e ferramental que possibilitará ao linguista de corpus conduzir os processos de formação de corpora de maneira guiada, vislumbrando antecipadamente etapas e complexidades, auxiliando no planejamento de pesquisas e consequentemente, na diminuição do esforço envolvido e no índice de retrabalho.

Na proposta do LAPELINC Framework, considera-se o atendimento ao requisito de fidelidade aos documentos originais para corpora históricos, incorporando a filosofia de fidedignidade em que se baseia o *workflow* do Método LAPELINC. No entanto, a extensão das funcionalidades a outros tipos de corpora que não só o de documentos históricos, possibilitou abranger outras variedades de iniciativas de pesquisa, sem renunciar à correta representação de documentos originais físicos.

Incluindo a possibilidade de associação dos textos a informações de geolocalização, o LAPELINC Framework possibilita a realização de estudos de geolinguística sobre os corpora, permitindo ainda a busca de informação sob critérios geográficos.

Com a possibilidade de se automatizar a formação de corpora a partir de textos oriundos de sítios da Web, o LAPELINC Framework possibilita a construção simplificada de corpora, possibilitando que diversos tipos de pesquisadores se debrucem sobre textos do gênero hipertextual sem uma sobrecarga de tempo e esforço desnecessários.

A inclusão de metainformações associadas diretamente à imagem, através da extensão de fac-símiles por meio de um processo de síntese de imagens, cria uma vinculação indissociável entre a imagem e o documento, de forma similar à fotografia estabelecida no Método LAPELINC (SANTOS; BRITO, 2017). A inclusão de um dicionário especializado de letras, grafias, abreviaturas e símbolos fornece ao pesquisador que esteja efetivando a transcrição de DDI ou fac-símiles, um auxílio considerável na busca pelo significado nos registros caligráficos.

Os módulos de anotação e edição incluídos disponibilizam funcionalidades imprescindíveis aos corpora linguísticos anotados, eliminando a necessidade de utilização de recursos externos, sem, contudo, eliminar o uso opcional de outra ferramenta. Ao incluir um módulo de anotação semântica, é possível associar aos textos através de seus recursos internos a compilação de estudos semânticos realizados, conforme uma teoria selecionada e disponível.

Com funcionalidades relacionadas a diferentes tipos de visualização dos textos, o framework permite ao pesquisador selecionar uma visão que mais se adeque à vertente teórica utilizada ou ao detalhamento requerido. Um historiador que necessite visualizar um dado documento poderá selecionar uma visualização em 3D da obra desejada, enquanto um linguista que esteja conduzindo estudos na área da sintaxe poderá visualizar o mesmo documento a partir de buscas sintáticas ou morfossintáticas que lhe retornem estruturas arbóreas apresentadas graficamente.

Finalmente, os recursos nativos de intercâmbio da informação permitem o reuso dos corpora criados por outros pesquisadores, estabelecendo diretrizes de trabalho que focam o

compartilhamento dos resultados e a continuidade dos trabalhos iniciados. Os benefícios poderão ser percebidos não apenas por pesquisadores que usufruam do uso de corpora construídos por terceiros, mas de forma limitada por toda a comunidade de linguística de corpus, que terá ao seu dispor uma metodologia que foque a continuidade e a contínua agregação de conhecimentos linguísticos aos fatos específicos de linguagem representados pelo corpus que utilize o LAPELINC Framework.

Por se tratar de uma metodologia em desenvolvimento, sua implementação através de recursos de software que a operacionalizem ocorrerá no decorrer da pesquisa, que visa apresentar à comunidade uma ferramenta apropriada e flexível para a construção de corpora linguísticos.

Referências

AHO, A. V. et al. **Compiladores:** princípios, técnicas e ferramentas. 2. ed. Porto Alegre: Pearson Education, 2007.

ALUÍSIO, S. M.; ALMEIDA, G. M. D. B. O que é e como se constrói um corpus? Lições aprendidas na compilação de vários corpora para pesquisa linguística. **Caleidoscópio**, Brasília, v. 4, p. 156-178, set-dez 2006.

ALVES, D. A. D. S.; MORINAKA, E. M. Procedimentos Metodológicos em Estudos da Tradução: Interface com as Linguísticas Sistêmico-Funcional e de Corpus. **Caderno das Letras**, Pelotas, v. 22, jan-jul 2014.

AUROUX, S. **A filosofia da linguagem**. Tradução de José Horta Nunes. 1ª. ed. Campinas: UNICAMP, 1998.

BERWANGER, A. R.; LEAL, J. E. F. **Noções de Paleografia e de Diplomática**. 5ª. ed. Santa Maria: Editora UFSM, 2015.

CARDOSO, S. A. **Geolinguística:** tradição e modernidade. São Paulo: Parágola Editorial, 2010.

COSTA, B. S. **Um framework integrado para a criação, o gerenciamento e a disponibilização de corpora digitais em língua portuguesa**. Projeto de Pesquisa de Doutorado (PPGLIN/UESB). Vitória da Conquista. 2019.

COSTA, B. S.; NAMIUTI, C.; SANTOS, J. V. Topografia de Documentos Digitais Imagem (DDI's) através da indexação de informações do método Lapelinc, Vitória da Conquista, 2019.

EVANS, D. Information about Corpus building and investigation: a on-line information pack about corpus investigation techniques for the Humanities. **Centre for Corpus Research**, University of Birmingham, 2008. Disponivel em:

<http://www.birmingham.ac.uk/documents/college-artslaw/corpus/intro/unit2.pdf>. Acesso em: 01 fev. 2019.

FAYAD, M. E.; SCHMIDT, D. C.; JOHNSON, R. E. **Building Application Frameworks:** object-oriented foundations of framework design. New York: Wiley, 1999.

FREITAS, C. Estudos linguísticos e Humanidades digitais: corpus e descorporificação. **Gragoatá**, Niterói, p. 1207-1227, setembro-dezembro 2017.

GALVES, C. O Corpus Tycho Brahe: Um corpus sintaticamente anotado do português histórico. **RBBA**, Vitória da Conquista, v. 8, p. 181-204, julho 2019. ISSN 23161205. Disponivel em: <https://periodicos2.uesb.br/index.php/rbba/issue/view/339>.

GUIMARÃES, E. **Semântica do acontecimento:** um estudo enunciatico da designação. 4ª. ed. Campinas: Pontes, 2017.

ISACA. **COBIT 5:** A Business Framework for the Governance and Management of Enterprise IT. EUA: ISACA, 2013.

MAGALHÃES, I. L.; PINHEIRO, W. B. **Gerenciamento de Serviços de TI na Prática:** uma abordagem com base na ITIL. São Paulo: Novatec, 2007.

MCENERY, T.; WILSON, A. **Corpus linguistics**. 2ª. ed. Edinburgh: Edinburgh University Press, 2001.

MENDES, A. **Linguística de Corpus e outros usos do corpus em Linguística. In:** Manual de Linguística Portuguesa. Berlin/Boston: Walter de Gruyter, 2016.

NAMIUTI, C. et al. Computação e Linguística: importante diálogo para pesquisas e preservação da memória nos novos meios das antigas fontes. **RBBA: Diálogo entre às ciências**, Vitória da Conquista, v. 2, p. 9-34, 2013. ISSN 2316-1205.

NAMIUTI, C.; SANTOS, J. V. De manuscritos históricos a corpora anotados: do Documento Digital Texto (DDT) ao corpus anotado. **Revista A Cor das Letras**, 17, 2016. 60-66.

PAIXÃO DE SOUSA, M. C. Memória do Texto. **Revista Texto Digital**, Santa Catarina, 2006.

PAIXÃO DE SOUSA, M. C. O Corpus Tycho Brahe: contribuições para as humanidades digitais no Brasil. **Filologia e Linguística Portuguesa**, São Paulo, v. 16, p. 53-93, dezembro 2014.

PORTELA, M. Humanidades digitais: as humanidades na era da Web 2.0, Número 38. out 2013. **Rua Larga - Revista da Reitoria da Universidade De Coimbra**, Coimbra, v. 38, 2013.

SANTOS, J. V.; BRITO, G. S. A Transposição de documentos manuscritos históricos jurídicos para o meio Digital através. In: _____ **E-Book do Congresso de**

Humanidades Digitais em Portugal: Construir pontes e quebrar. Lisboa: Universidade Nova de Lisboa, 2017.

SANTOS, J. V.; NAMIUTE, C. O objeto livro: a complexidade da forma e o digital. **Anais do X Congresso Internacional da ABRALIN**, Rio de Janeiro, 2017.

SANTOS, J. V.; NAMIUTI, C. DE MANUSCRITOS HISTÓRICOS A CORPORA ANOTADOS: DO DOCUMENTO FISICO (DF) AO DOCUMENTO DIGITAL IMAGEM (DDI). **A Cor das Letras**, Feira de Santana, 2016.

TEUBERT, W. Editorial. **International Journal of Corpus Linguistics**, Amsterdam, v. 1, n. 1, 1996. ISSN 1384-6655.