ISSN 1415 - 8973

eASSIGNER: UMA PROPOSTA DE AUTOMAÇÃO DAS EDIÇÕES DE ANOTAÇÕES XML DO eDICTOR

eASSIGNER: A Proposal for automating the eDICTOR XML Annotations Editions

Cristiane Namiuti Temponi*
Universidade Estadual de Campinas, Campinas, São Paulo, Brasil

Pablo Picasso Feliciano de Faria**
Universidade Estadual do Sudoeste da Bahia, Vitória da Conquista, Bahia, Brasil

Luiz Fernando Cardeal de Souza***
Instituto Federal da Bahia, Vitória da Conquista, Bahia, Brasil

Resumo: Este trabalho discute a importância da aplicação das técnicas de Aprendizado de Máquina - ligadas à Inteligência Artificial (IA) - como ferramentas de auxílio aos trabalhos dos linguistas, particularmente no que se refere à Linguística de *Corpus*. Para atingir essa proposta, o trabalho define a Linguística de *Corpus*, a Linguística Computacional e descreve utilizações atuais da IA enfatizando os problemas relacionados à Linguística e as estratégias modernas em busca de soluções. A partir daí, descreve a importância do uso de *softwares* de anotação em *corpora* eletrônicos e a decorrente necessidade de automatizar algumas dessas operações através do desenvolvimento do *software* eAssigner, apresentando as suas características, limitações e estágio de desenvolvimento. Para ilustrar a necessidade de uso do *software* são apresentados resultados de alguns testes realizados em amostra de documentos do *Corpus* DoViC.

Palavras chave: Anotação. Aprendizado de Máquina. Corpora. Linguística Computacional. Linguística de *Corpus*.

Abstract: This paper discusses the importance of the application of Machine Learning techniques - linked to Artificial Intelligence (AI) - as tools to aid the work of linguists, particularly with regard to Corpus Linguistics. In order to reach this proposal, the work defines Corpus Linguistics, Computational Linguistics and describes current uses of the IA emphasizing the problems related to Linguistics and the modern strategies in search of solutions. From there, it describes the importance of the use of electronic annotation software in corpora and the resulting need to automate some of these operations through the development of the eAssigner software, presenting its characteristics, limitations and stage of development. To illustrate the need to use the software are presented results of some tests performed in some DoViC Corpus documents.

Keywords: Annotation. Machine Learning. Corpora. Computational Linguistics. Corpus Linguistics.

1 INTRODUÇÃO

Um dos problemas de grande interesse para os estudos em linguística histórica está ligado ao manuseio e tratamento de *corpora*: trata-se da necessidade de construir bibliotecas de documentos com anotações conforme Faria e Galves (2016) explicam:

* Doutora em Linguística (UNICAMP); Professora do Departamento de Estudos Linguísticos e Literários da Universidade Estadual do Sudoeste da Bahia (UESB); Professora do PPGLIN (Programa de Pós-Graduação em Linguística). Coordenadora, juntamente com o Professor Dr. Jorge Viana Santos, do Laboratório de Pesquisa em Linguística de Corpus (LAPELINC). Pesquisadora da Fundação de Amparo à Pesquisa do Estado da Bahia (FAPESB) e do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). E-mail: cristianenamiuti@uesb.edu.br.

** Doutor em Linguística (UNICAMP), com ênfase em Aquisição de Linguagem e Linguística Computacional. Professor do Departamento de Linguística da Universidade Estadual de Campinas. Tem interesse nas áreas de formalismos gramaticais, aquisição de linguagem, teorias formais de aprendizagem, processamento automático (parsing, tradução, etc.) e psicolinguística. Possui formação interdisciplinar, que inclui o bacharelado em Ciências da Computação, pós-graduação (aperfeiçoamento) em Sistemas de Informação, mestrado e doutorado em Linguística. E-mail: pablofaria@gmail.com.

*** Mestrando junto ao Programa de Pós-Graduação em Linguística (PPGLin) da Universidade Estadual do Sudoeste da Bahia (UESB) e Bacharel em Processamento de Dados pela Universidade Federal da Bahia (UFBA). Professor do Instituto Federal da Bahia (IFBA) Campus de Vitória da Conquista – BA. E-mail: fcardeal@gmail.com.

ISSN 1415 - 8973

> *Corpora* anotados são importantes em todos os ramos da linguística, uma vez que constituem bases de dados perenes sobre as quais se podem efetuar análises qualitativas e quantitativas de vários tipos, que complementam outras abordagens como o recurso à intuição dos falantes ou ainda estudos baseados em experimentos, prática corrente em aquisição da linguagem e cada vez mais em análises sintáticas. Em linguística histórica, uma vez que não há falantes nativos disponíveis, os *corpora* são indispensáveis. (FARIA; GALVES, 2016, p. 303).

A construção desses *corpora* anotados deve ser feita de forma a torná-los acessíveis através de programas de busca e, para tanto, investe-se na tarefa de transcrição e edição das fontes documentais. Faria e Galves (2016, p.302), enfatizam que "[...] Esta construção, atualmente, é semiautomática e envolve, geralmente, programas de computador que implementam algoritmos de aprendizagem (de máquina)".

O processo de construção de corpora eletrônicos mediante a edição de textos antigos envolve várias operações como: junção ou segmentação de termos, modernização de grafia, expansão, correção etc. Como cada operação dessas ocorre várias vezes ao longo de um texto, o processo tende a se tornar repetitivo e sujeito a erros. Além disso, operações feitas em um documento possivelmente terão que ser repetidas em outros documentos do mesmo *corpus*, o que já justifica a necessidade de ferramentas computacionais para auxílio dos pesquisadores.

O artigo está organizado da seguinte forma: na seção 2 é apresentada uma introdução à Linguística Computacional, enfatizando as suas áreas de atuação e apresentado o conceito de aprendizado (ou aprendizagem) de máquina. Na seção 3 é apresentado o processo de produção de *corpora* para a pesquisa em Humanidades Digitais, enfatizando o uso de etiquetas XML; também apresenta o software eDictor, comentando suas potencialidades e limitações. Na seção 4, é apresentado o projeto de software eAssigner deverá complementar o eDictor na atividade de anotação de *corpora* através da automatização de anotações XML. A seção 5 apresenta as considerações finais sobre o trabalho e perspectivas de desenvolvimento futuro.

2 LINGUÍSTICA COMPUTACIONAL

Conforme Othero e Menuzzi (2005, p. 22), "a área responsável pela investigação do tratamento computacional da linguagem e das línguas naturais é conhecida como 'Linguística Computacional'". Então, fica claro que um dos objetos de estudo da Linguística Computacional é a comunicação humano-computador. Outros ramos de investigação da área são: a Análise Linguística automática (morfologia, sintaxe e semântica), a modelagem de processos ligados à linguagem, como a aquisição da linguagem, entre outros. Vieira e Lima assim definem:

> A Linguística Computacional pode ser entendida como a área de conhecimento que explora as relações entre linguística e informática, tornando possível a construção de sistemas com capacidade de reconhecer e produzir informação apresentada em linguagem natural. (VIEIRA e LIMA, 2001, p.1).

Segundo Othero e Menuzzi (2005, p.22), a Linguística Computacional pode ser dividida em duas subáreas: a **linguística de *corpus*** e o **processamento de linguagem natural (PLN)**. Como muitos trabalhos envolvem as duas áreas, nem sempre ocorre uma divisão nítida entre elas, ainda conforme os autores citados.

ISSN 1415 - 8973

3 LINGUÍSTICA DE CORPUS

Conforme Othero e Menuzzi (2005, p.22), "[...] a Linguística de Corpus preocupa-se basicamente com o trabalho a partir de *corpora* eletrônicos que contenham amostras de linguagem natural". De acordo com Berber Sardinha:

> A Linguística de *Corpus* se ocupa da coleta e exploração de *corpora*, ou conjuntos de dados linguísticos textuais que foram coletados criteriosamente com o propósito de servirem para a pesquisa de uma língua ou variedade linguística. Como tal, dedica-se à exploração da linguagem através de evidências empíricas, extraídas por meio de computador. (BERBER SARDINHA, 2000, p. 2).

O autor ainda lembra que antes de existirem computadores, já existiam *corpora* e cita o *Corpus* Helenístico criado por Alexandre o Grande na Grécia Antiga. Outro exemplo é a criação de *corpora* na Antiguidade e na Idade Média com citações da Bíblia.

Othero e Menuzzi (2005, p.23) afirmam que nem todos os trabalhos com linguística de *corpus* estão associados ao desenvolvimento de algum *software*, podem estar voltados ao estudo de fenômenos linguísticos e sua ocorrência em grandes amostras de uma determinada língua.

4 PROCESSAMENTO DE LINGUAGEM NATURAL (PLN)

A área de PLN, de acordo com Othero e Menuzzi tem a seguinte preocupação:

> A PLN preocupa-se diretamente com o estudo da linguagem voltado para a construção de *softwares* e sistemas computacionais específicos, como tradutores automáticos, *chatterbots*[1], *parsers*[2], reconhecedores automáticos de voz, geradores automáticos de resumos etc. (OTHERO; MENUZZI, 2005, p. 23).

Assim, fica bem claro que a PLN é a área da Linguística Computacional focada no desenvolvimento de programas capazes interpretar e/ou de gerar informações em Língua Natural. Esses programas devem ser capazes de analisar sentenças e, conforme Faria e Galves (2016):

> A análise automática de uma sentença consiste em atribuir uma ou mais estruturas sintáticas a ela, de modo que as relações entre as palavras sejam explicitadas seja pela delimitação dos constituintes sintáticos que elas formam, seja pela identificação da função sintática dos elementos. Os analisadores modernos são, de modo geral, total ou parcialmente probabilísticos, isto é, produzem análises possíveis (quando mais que uma) de uma sentença e as ordenam conforme a probabilidade de cada uma. A análise com maior probabilidade é geralmente tida como a "melhor" (i.e., provavelmente mais correta) análise para uma dada sentença (FARIA; GALVES, 2016, p. 306).

Dessa forma, um programa analisador será capaz de aprender a executar a análise a partir de análises feitas manualmente por linguistas. De acordo com Faria e Galves (2016, p. 306), "a 'aprendizagem' de um analisador consiste, portanto, em receber exemplos de análise

[1] *Chatterbot* é um programa de conversação automática, um robô de conversação, tipicamente usado em atendimentos eletrônicos (Nota dos Autores).

[2] *Parser* é um programa que faz a análise sintática de um texto. Pode ter vários objetivos: ajudar na tradução, marcar textos etc.

ISSN 1415 - 8973

e construir um modelo probabilístico que lhe permita aplicar análises sobre novas sentenças, inclusive aquelas estritamente inéditas".

Coppin (2010, p. 233) afirma que o Aprendizado de Máquina é um dos mais importantes segmentos da Inteligência Artificial. Assim define o aprendizado por treinamento: "é a capacidade de uma máquina ou sistema aprender a classificar entradas de acordo com um conjunto finito de classificações". Dessa forma, considerando a necessidade dos linguistas, são feitas várias anotações do mesmo tipo e são submetidas ao *software*. Este "lê" as anotações, executa um mecanismo de aprendizado que pode ser reconhecimento de padrões e, a seguir, o *software* poderá ser capaz de reconhecer a ocorrência desses padrões.

Russel e Novig (2013, p.383) descrevem essa forma de aprendizado como "aprendizagem a partir de exemplos", também conhecida como aprendizagem supervisionada. Um agente inteligente (programa), analisa vários exemplos, registra esse conhecimento e passa a ser capaz de repeti-lo.

5 PRODUÇÃO DE *COPRPORA* PARA A PESQUISA EM HUMANIDADES EM AMBIENTE DIGITAL

A fidelidade e a confiabilidade dos *corpora* eletrônicos em relação aos documentos originais são alguns dos pilares de sustentação dos estudos linguísticos, conforme enfatizam Namiuti-Temponi, Viana Santos e Leite, retomando Paixão de Sousa (2006):

> [...]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 linguístico, em qualquer quadro teórico, deve pressupor. No caso dos *corpora* eletrônicos, esse pressuposto fundamental precisa ser integrado com requerimentos impostos pela vertente computacional e linguística 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-TEMPONI, VIANA SANTOS; LEITE, 2011, p. 2833).

Para os autores, é necessário produzir os *corpora* de forma a facilitar o tratamento computacional, mas o processo deve ser feito de tal maneira que seja mantida a forma original de cada documento. O uso da linguagem XML[3] é uma estratégia que vem sendo adotada por linguistas na criação de *Corpora* como o *corpus* Documentos Oitocentistas de Vitória da Conquista (DOViC).

6 XML E ETIQUETAS DE MARCAÇÃO

A linguagem XML utiliza etiquetas (*tags* em inglês) para delimitar e identificar as marcações. Uma das características dessa linguagem é que as etiquetas são livres, podem ser criadas de acordo com a necessidade de utilização. Por isso a linguagem é dita extensível. Portanto, um documento XML é um documento de texto que representa dados de maneira estruturada, utilizando um conjunto de marcadores.

Abaixo exibimos um exemplo de uma edição XML segundo os padrões do Corpus Tycho Brahe, mostrando os marcadores utilizados em uma suposta anotação do tipo "junção" onde a palavra original era "sepa rado" e, após a junção, é mantida a palavra original e é feita a operação de junção, mostrando a palavra corrigida:

[3] XML – eXtensible Markup Language – Ou linguagem de marcadores extensíveis (tradução livre).

ISSN 1415 - 8973

<w>	Indica o início de um bloco
<o>sepa rado</o>	Indica o texto original
<e t="jun">separado</e>	Indica a junção e mostra a palavra corrigida
</w>	Fim do bloco

Caracterizando o uso de etiquetas, Galves e Britto (1999, p.1) definem requisitos para um sistema de anotação morfológica:

> Adequação descritiva: as etiquetas têm que representar e discriminar adequadamente as categorias necessárias à descrição dos enunciados presentes na língua em geral.
> Recuperabilidade da informação: O objetivo do corpus anotado é permitir aos estudiosos da história do português obterem de maneira rápida e confiável as informações necessárias para desenvolver análises sincrônicas ou diacrônicas de aspectos lexicais, morfológicos e sobretudo sintáticos da língua. O conjunto de etiquetas tem que ser construído e aplicado de modo a permitir recuperar da maneira mais econômica e exaustiva possível essas informações.
> Simplicidade computacional: O número total de etiquetas diferentes compondo o conjunto deve ser compatível com um tratamento computacional do corpus, nomeadamente com o treinamento de um etiquetador automático aplicado a este. (GALVES; BRITTO, 1999, p.1).

Esse processo é feito com o auxílio de um tipo de ferramenta computacional chamada "Ferramenta de Anotação".

Faria, Paixão de Sousa e Kepler (2010, p.2), identificaram e comentaram diversas ferramentas de anotação disponíveis na época que, apesar de possuírem muitos recursos, prosseguem os autores, não lidam de forma adequada com os diversos níveis de edição para os marcadores dos textos. Isso ensejou o desenvolvimento de uma ferramenta própria, o eDictor.

6.1 O EDICTOR

O eDictor (anteriormente E-Dictor) é uma ferramenta concebida para auxiliar a edição eletrônica em XML de textos antigos para fins de análise linguística automática, de acordo com Paixão de Sousa, Kepler e Faria (2009, p.1). O eDictor permite operar com vários níveis de edição: Junção, Segmentação, Grafia, Modernização, Expansão, Correção, Pontuação (e é possível criar novos níveis, de acordo com a necessidade do pesquisador), ainda segundo os autores. Outra característica importante é possibilidade de visualizar o texto original e o texto anotado (editado).

De acordo com Paixão de Sousa, Kepler e Faria, o eDictor apresenta as seguintes possibilidades de uso:

> Flexibilidade dos formatos gerados, permitindo tanto a leitura humana como a leitura automática; Garantia da qualidade filológica da edição por se tratar de um editor especializado; Software livre: o que dá a possibilidade de trabalho no código-fonte (alterar o programa original para atender a necessidades específicas); Previsão de continuidade do programa; Transferibilidade garantida. Ferramenta completa: O resultado combina correção do reconhecimento e edição de variação de grafia (PAIXÃO DE SOUSA; KEPLER; FARIA, 2009, p.22).

De acordo com Paixão de Sousa (2011, p. 21) o eDictor tem uma limitação importante: não é "treinável", o que tem duas implicações:

- Os resultados das anotações não são transferidos para o restante do acervo;
- Os resultados das anotações não são transferidos para outros projetos e acervos.

ISSN 1415 - 8973

Conforme já foi mencionado em outra seção, isso implica em retrabalho, aumentando a possibilidade de erros de anotação. Para superar as limitações apontadas, está em fase de desenvolvimento o *software* eAssigner.

6.2 O eASSIGNER

O eAssigner é um *software* em desenvolvimento, que utilizará os algoritmos de Aprendizado de Máquina para analisar as anotações XML feitas em *Corpora* digitais para aprender e replicar essas anotações no mesmo *corpus* ou propagar em outros *corpora*. A principal vantagem é superar a limitação apontada no eDictor.

O eAssigner está sendo desenvolvido para trabalhar em conjunto com o eDictor, compondo um conjunto de ferramentas de auxílio aos pesquisadores, reduzindo o esforço humano e aumentando a precisão dos resultados obtidos na análise de *Corpora*.

6.3 ILUSTRAÇÃO DA NECESSIDADE DO eASSIGNER

Para ilustrar a necessidade de uso do *software*, os autores utilizaram o *Corpus* DoVic, de onde foram extraídos e estudados oito documentos. Todos são cartas de alforria emitidas em favor de escravos. Os documentos foram transcritos segundo o método LAPELINC descrito em (NAMIUTI; SANTOS, 2016). Em seguida foram editados e revisados por linguistas utilizando o *software* eDictor. Os seguintes resultados foram observados:

No documento 1 pode-se observar que o nome "Jose" foi modernizado para "José" na edição número 13, mas não foi modernizado na edição número 23 do mesmo documento, conforme demonstrado a seguir:

```
<w id="13">
<o>Jose</o>
<e t="mod">José</e>                Feita uma MODERNIZAÇÃO
<m v="NPR"/>
</w>

.....

<w id="23">
<o>Jose</o>
<m v="NPR"/>
</w>
```

Observem o que acontece ao observar vários documentos diferentes:

Documento 1

```
<w id="28">
<o>assi-<bk t="l" id="bk_3"/> gnado</o>
<e t="jun">assi-gnado</e>
<e t="gra">assignado</e>
<e t="mod">assinado</e>
<m v="VB-AN"/>
</w>
```

Documento 3

```
<w id="35">
<o>assignado</o>
```

ISSN 1415 - 8973

</w>

Documento 4
<w id="275">
<o>assignado</o>
<e t="mod">**assinado**</e>
<m v="VB-AN"/>
</w>

Documento 6
<w id="310">
<o>assignado</o>
<e t="mod">**assinado**</e>
<m v="VB-AN"/>
</w>

A palavra "assignado" teve a sua grafia modernizada manualmente para "assinado" em três documentos diferentes (1, 4 e 6), dentro da nossa amostra de oito documentos. Observe também que no documento 3 a grafia não foi modernizada. Em uma pequena amostra do *corpus*, já é possível perceber a importância do desenvolvimento do eAssigner. E há ainda uma outra situação a ressaltar: o eAssigner terá que lidar com casos de ambiguidade, em que uma mesma forma arcaica pode ter duas ou mais formas modernizadas possíveis, a depender do contexto de ocorrência. Digamos, como exemplo hipotético, que se encontre a forma "õde" cuja modernização seja, em alguns casos, "onde" e noutros casos "aonde". Neste caso, é preciso levar o contexto em consideração para determinar qual a edição mais provável.

6.4 ESTRUTURA DO eASSIGNER

De forma similar ao eDictor, será capaz de ler textos em formato simples (XML) e atualizar esses mesmos textos.
O *Software* tem dois modos operacionais:

- Modo Treinamento: para ler as marcações, aprender e registrar na Base de Conhecimento (veja módulos adiante);
- Modo Operação: para ler e atualizar um *corpus* de acordo com as marcações e a base de conhecimento.

A sua estrutura essencial é composta dos seguintes módulos:

- Base de Conhecimento: é onde o eAssigner registrará as anotações que aprender. Inicialmente é um arquivo vazio e, à medida que o *software* aprende, registrará informações como: o padrão a reconhecer; a operação a executar; frequência de ocorrência (para utilização de métodos estatísticos de aprendizado e para gerar relatórios posteriormente);
- Módulo de Carga: carrega para a memória a Base de Conhecimento para tornar o processo mais ágil;
- Módulo de Leitura: permite reconhecer um *corpus*, abrir e ler o seu conteúdo, passando os dados para o módulo *Tokenizer*;
- Módulo *Tokenizer*: separa em blocos as tags XML dos dados lidos para permitir a análise de padrões. Cada tag identificada é chamada de *token*;
- Módulo Analisador de Padrões: recebe os *tokens* e verifica similaridade na Base de Conhecimento. Se estiver no **Modo Treinamento**, dá ao operador a oportunidade de indicar se é para aprender essa marcação. Se estiver no **Modo**

ISSN 1415 - 8973

Operação e o *token* existir na base de conhecimento, ativa o Módulo de Atualização.

- Módulo de Atualização: repete no *corpus* a marcação correspondente indicada na Base de Conhecimento.
- Módulo de Manutenção da Base de Conhecimento: atualiza a base de conhecimento à medida que aprende novas marcações. Trabalha em conjunto com o Analisador de Padrões.
- Módulo de Log: registra todas as operações executadas, permitindo eventuais recuperações de erro ou identificação de operações executadas.

6.4 ESTADO DO DESENVOLVIMENTO

Até a presente data, estão operacionais os módulos de Carga, Leitura, Tokenizer, Manutenção da base de conhecimento e Log. Os demais módulos estão em fase de desenvolvimento e testes.

7 CONSIDERAÇÕES FINAIS

Ao longo deste trabalho, foram apresentados alguns importantes aspectos para a construção de *Corpora* digitais, suas características, e fatores que tornam o processo mais lento ou impreciso. Foram definidos os conceitos de Linguística de Corpus e Linguística Computacional e suas características foram apresentadas. A partir daí, foi feita uma breve apresentação da Aprendizagem de Máquina e mostrada a sua relação e a sua importância para a Linguística.

Uma vez visitada essa importante base teórica, foi descrita a utilização, importância, características e limitações do software eDictor, um software nacional, desenvolvido por pesquisadores da UNICAMP e USP, mostrando os avanços da Linguística de Corpus desenvolvida no Brasil. A partir dessas conhecidas limitações que foram, inclusive, apontadas pela própria equipe de desenvolvimento do eDictor, foi projetado o software eAssigner que objetiva complementar as operações já disponíveis. Esse é mais um trabalho desenvolvido por brasileiros e que, assim como o eDictor, utiliza ferramentas livres, tornando-o independente de fabricantes e de equipamentos; poderá ser modificado, aperfeiçoado por outras equipes, viabilizando o uso coletivo e a adequação a outras necessidades que venham a ser identificadas no futuro.

Espera-se otimizar o trabalho dos linguistas na construção de *Corpora* eletrônicos com o aumento de velocidade e de precisão. No momento, não é possível quantificar, mas, considerando o exemplo real ilustrado na subseção 4.1, certamente o ganho será significativo.

REFERÊNCIAS

BERBER SARDINHA, Tony. *O que é um corpus representativo?* São Paulo: LAEL PUCSP, 2000. Disponível em http://www2.lael.pucsp.br/direct/DirectPapers44.pdf acesso em 11 out. 2015.

COPPIN, Ben. *Inteligência Artificial*. Rio de Janeiro: LTC, 2010.

DIAS-DA-SILVA, Bento Carlos. O estudo Linguístico-Computacional da Linguagem. *Letras de Hoje*. Porto Alegre. v. 41, nº 2, p. 103-138, junho, 2006. Disponível em

ISSN 1415 - 8973

<http://revistaseletronicas.pucrs.br/ojs/index.php/fale/article/viewFile/597/428> acesso em 13 out. 2015.

FARIA, Pablo; GALVES, Charlotte. Criando "Bancos de Árvores": O Sistema de Anotação e o Processo Automático. *Cadernos de Estudos Linguísticos*. Campinas: v. 58, n. 2 p. 299-315, maio/ago./2016. Disponível em http://revistas.iel.unicamp.br/index.php/cel/article/view/5133. Acesso em 30 dez. 2016.

FARIA, Pablo P. F.; PAIXÃO DE SOUSA, M. C.; KEPLER, F. N. *An Integrated Tool for Annotating Historical Corpora.* The Fourth Linguistic Annotation Workshop (LAW IV) at The 48th Annual Meeting of the Association for Computational Linguistics (ALC 2010), Uppsala, 2010. (Congresso).

GALVES, Charlotte; BRITTO, Helena. *A Construção do Corpus Anotado do Português Histórico Tycho Brahe:* o sistema de anotação morfológica. USP. São Paulo: 1999. Disponível em https://www.ime.usp.br/~tycho/participants/c_galves/galves_e_britto.htm. Acesso em 13 out. 2015.

LUGER, George F. *Inteligência Artificial***.** 6ª ed. Tradução de Daniel Vieira. São Paulo: Pearson Education do Brasil, 2013.

NAMIUTI, Cristiane; SANTOS, Jorge Viana. Novos desafios para antigas fontes: a experiência DOViC na nova linguística histórica. In.: *E-Book do Congresso de Humanidades Digitais em Portugal: Construir pontes e quebrar barreiras na era digital – 2015*. Lisboa: Universidade Nova de Lisboa, 2016a (no prelo).

NAMIUTI, Cristiane; COSTA, Aline Silva. Reflexões sobre anotação sintática e ferramentas de busca - Uso da linguagem XML para anotação sintática no corpus digital DOViC. *Letras & Letras*. Uberlândia, v.30, n.2, 2014, p. 82-103. Disponível em http://www.seer.ufu.br/index.php/letraseletras/article/view/27855/15804. Acesso em 2 ago. 2016.

NAMIUTI, Cristiane; VIANA SANTOS, Jorge; LEITE, Cândida Mara Brito. *Propostas e Desafios dos Novos Meios das Antigas Fontes:* a preservação da memória pela Linguística de *Corpus*. Trabalho apresentado no IX Colóquio do Museu Pedagógico. UESB, Vitória da Conquista: 2011. Disponível em http://periodicos.uesb.br/index.php/cmp/article/viewFile/2717/2382. Acesso em 2 ago. 2016.

OTHERO, Gabriel de Ávila; MENUZZI, Sérgio de Moura. *Linguística Computacional:* teoria & prática. São Paulo: Parábola Editorial, 2005.

OTHERO, Gabriel de Ávila. Linguística Computacional: uma breve introdução. *Letras de Hoje*. Porto Alegre. v. 41, nº2, p. 341-351, junho, 2006. <http://revistaseletronicas.pucrs.br/ojs/index.php/fale/article > acesso em 13 out. 2015.

PAIXÃO DE SOUSA, M. C. Memórias do Texto. In: *Revista Texto Digital*, ISSN 1807-9288, ano 2 n.1 2006a.<http://www.textodigital.ufsc.br/num02/paixao.htm>

ISSN 1415 - 8973

PAIXÃO DE SOUSA, M. C. *A anotação semiautomática de divergências de grafia como fundamento para o processamento automático de textos antigos*: Uma experiência na Brasiliana Digital. 18º Intercâmbio de Pesquisas em Linguística Aplicada, PUC, São Paulo, 2011.

PAIXÃO DE SOUSA, Maria Clara; KEPLER, Fabio Natanael; FARIA, Pablo Picasso Feliciano de. *e-Dictor*. Versão 1.0 beta 10, 2013. Programa de Computador. Disponível em: http://edictor.net/download. Acesso em 01 jun.2016. Acesso em 15 maio 2016.

_____. *eDictor*: Novas perspectivas na codificação e edição de corpora de textos históricos. In: VIII Encontro de Linguística de Corpus, 2009, Rio de Janeiro. Resumos, 2009.

RUSSELL, Stuart; NOVIG, Peter. *Inteligência Artificial*. 3ª ed. Rio de Janeiro: Elsevier, 2013 [e-book].

SANTOS, Jorge Viana. *Técnicas de transporte do texto manuscrito para o meio digital*. Trabalho apresentado na I Oficina de Linguística de Corpus da Bahia (UEFS, UESB, UFBA). Feira de Santana, Brasil, Dezembro 15-17, 2010.

VIEIRA, Renata; LIMA, V.L.S. *Linguística Computacional:* princípios e aplicações. In: IX Escola de Informática da SBC-Sul. Luciana Nedel (Ed.) Passo Fundo, Maringá, São José. SBC-Sul, 2001.