IX Encontro de Linguística de Corpus
Porto Alegre – PUCRS – 8 e 9 de outubro de 2010

AELIUS: UMA FERRAMENTA PARA ANOTAÇÃO AUTOMÁTICA DE CORPORA USANDO O NLTK

Leonel Figueiredo de Alencar (UFC)

E-Mail: leonel D O T de D O T alencar A T ufc D O T br
Homepage: **http://www.leonel.profusehost.net/**

INTRODUÇÃO

Este trabalho descreve aspectos do Aelius, pacote em Python que, utilizando a biblioteca Natural Language Toolkit (NLTK) (Bird, Klein e Loper, 2009), se destina ao pré-processamento de textos, construção de etiquetador morfossintático, anotação de corpora e auxílio de revisão humana de anotação automática. O Aelius não se limita a oferecer uma interface simplificada para algumas dessas funcionalidades no NLTK, mas o complementa de várias formas, com destaque para o aumento da eficácia da etiquetagem morfossintática por meio da utilização de um algoritmo próprio para lidar com palavras de inicial maiúscula.

Os módulos de construção de etiquetador e anotação morfossintática de corpora são exemplificados por meio de sua aplicação, no âmbito do projeto CORPTEXLIT (Alencar, 2010), a textos do português do séc. XIX ainda não contemplados pelo Corpus Histórico do Português Tycho Brahe (CHPTB). O nível alcançado de acurácia da etiquetagem supera o de ferramentas análogas livremente disponíveis, voltadas sobretudo para o português contemporâneo.

O Aelius permite preencher, portanto, lacunas na lingüística de corpus do português de orientação diacrônica, ao mesmo tempo em que contribui para sanar a escassez de corpora e *language models* do português no NLTK e acrescenta a essa biblioteca algumas funções bastante úteis para o desenvolvimento de etiquetadores mais eficazes.

OBJETIVOS

Visamos, inicialmente, a aproximar da lingüística computacional, valendo-se da amigabilidade do NLTK, a comunidade brasileira de Letras e Lingüística, o que pressupõe a disponibilização de ferramentas ainda mais amigáveis, *language models* e mais corpora anotados representativos da diversidade do português. Por outro lado, a lingüística histórica do português, apesar da cobertura do CHPTB para o período de 1500 a 1800, ainda não dispõe de uma quantidade suficiente de textos anotados do século XIX.

Com este trabalho, pretendemos contribuir para preencher essas lacunas. Para tanto, desenvolvemos o Aelius, que, recorrendo não ao apenas ao NLTK, mas a algoritmos próprios, complementa essa biblioteca de várias formas, na execução das seguintes tarefas:

(i) Pré-processamento de corpora
(ii) Construção de *language models* e etiquetadores com base num corpus anotado
(iii) Avaliação do desempenho de um etiquetador
(iv) Comparação entre diferentes anotações de um texto

O impulso para construção do Aelius foi dado pela necessidade de dispor de um anotador morfossintático de alta acurácia para textos literários do século XIX, objetivando um índice inicial de 95% na anotação de textos desse tipo, pois isso já permite simplificar imensamente a revisão humana, base, por sua vez, no processo de *bootstraping*, para versões mais robustas do etiquetador (Naumann, 2004; Ule e Hinrichs, 2004).

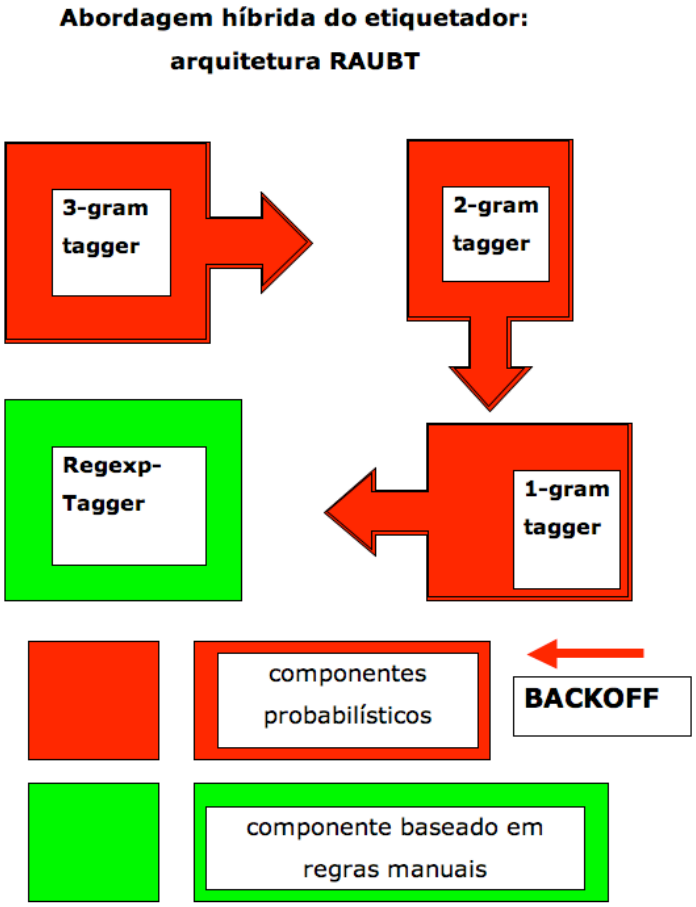

Figura 1: Arquitetura do etiquetador

METODOLOGIA

Seguindo Bird, Klein e Loper (2009, p. 198-208), o etiquetador do Aelius adota uma abordagem *híbrida* (Voutilainen, 2004) que recorre a regras, formuladas manualmente em expressões regulares, para etiquetar as palavras inexistentes no *language model*. O sistema adota a estratégia de *backoff* sugerida por esses autores, encadeando etiquetadores estocásticos baseados em n-gramas (onde 1 <= n <= 3), em ordem decrescente, a um etiquetador baseado em expressões regulares, conforme proposta de Perkins (2010), configurando a arquitetura que denomina RAUBT (ver Figura 1).

O Aelius comporta uma série de funções em Python que pré-processam o corpus de treino conforme vários parâmetros especificados pelo usuário. Diferentemente das classes do NLTK usadas para construção de etiquetadores baseados em n-gramas, o usuário, no Aelius, pode especificar um conjunto de etiquetas a serem ignoradas na construção do modelo bem como dividir o corpus de base em um número específico de blocos e embaralhá-los de forma aleatória, o que permite obter um corpus de treino e um corpus de teste mais balanceados (Kepler, 2005).

No desenvolvimento do etiquetador, procedemos incrementalmente, repetindo várias vezes o ciclo editar – compilar – testar – depurar (ver Figura 2). Esse processo envolveu não só a otimização do etiquetador baseado em expressões regulares, mas também a eliminação, do CHPTB, de tokens com etiquetas que não se referem à análise

morfossintática bem como correções de inconsistências do próprio CHPTB. A grande quantidade de erros relacionadas à etiqueta NPR (nome próprio) levou-nos a desenvolver e implementar em Python um algoritmo para distinguir nomes próprios de outras palavras com inicial maiúscula. Esse algoritmo supera grave deficiência do etiquetador de expressões regulares do NLTK, que não leva em conta o contexto onde um token ocorre.

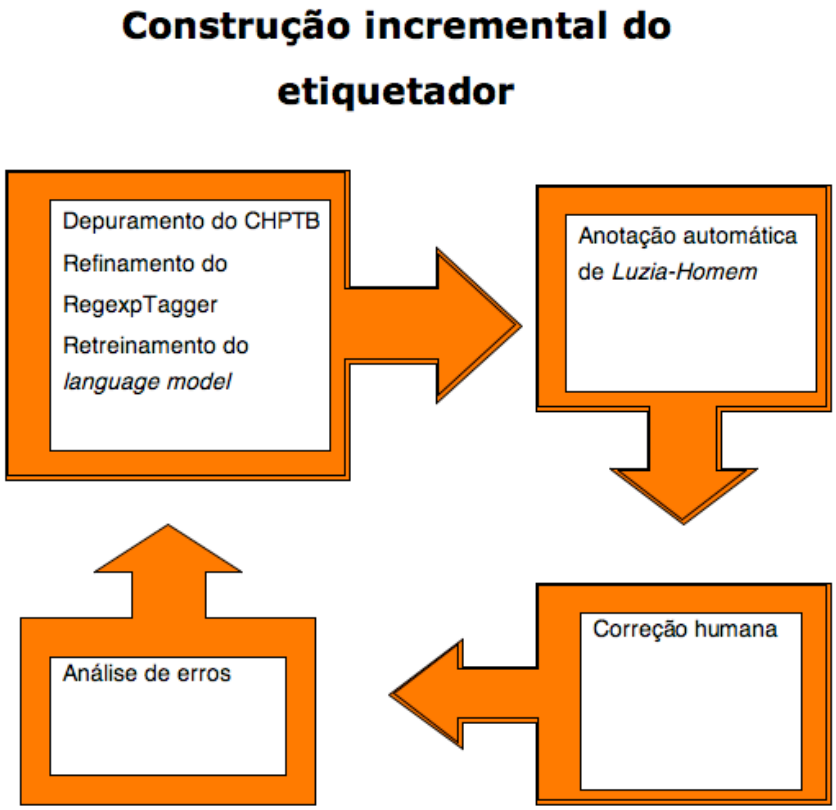

Figura 2: Etapas da construção do etiquetador

RESULTADOS

As otimizações realizadas ao longo do ciclo da Figura 2 permitiram obter melhores resultados na aplicação do etiquetador tanto aos corpora de teste quanto ao romance *Luzia-Homem* (1903), de Domingos Olímpio (1850-1906). No primeiro caso, alcançamos um índice de acurácia médio de 94,21% em 5 execuções da aplicação do etiquetador treinado em 75% das 67141 sentenças do CHPTB (totalizando 1431475 tokens etiquetados), divididas em 1000 blocos aleatoriamente embaralhados, nos 25% restantes. Esse índice está mais de 4% acima dos relatados por Perkins (2010) em experimentos com o NLTK na etiquetagem de corpora do inglês usando a mesma arquitetura RAUBT da Figura 1. Embora essa língua aparentemente seja mais ambígua que o português, o conjunto de etiquetas do CHPTB é mais complexo, complexidade essa responsável por muitas ambigüidades. Esse nível de precisão, alcançado pela exploração sistemática das regularidades da morfologia flexional e derivacional do

português, é também bem próximo dos relatados, para a língua portuguesa, por abordagens puramente estatísticas, como Kepler (2005) e Lácio-Web (s.d.), os quais oscilam em torno de 95% para a maioria dos textos. A acurácia do Aelius é equiparável à obtida por Garcia e Gamallo (2010) para o português europeu, no âmbito do FreeLing, projeto de software livre análogo ao NLTK (mas direcionado ao desenvolvedor de tecnologia da linguagem natural e não ao usuário final).

Na etiquetagem dos quatro primeiros capítulos de *Luzia-Homem* (totalizando 7404 tokens, cerca de 10% do romance), obtivemos, com uma versão do etiquetador treinada em 100% das sentenças de uma versão do CHPTB depurada de erros e inconsistências de anotação, um total 95.08% de acertos. Etiquetadores treinados pelo VLMMTagger na versão original do CHPTB e nessa mesma versão depurada, contudo, tiveram desempenho inferior, atingindo, respectivamente, 93.54% e 94.41%. Cremos que a vantagem do Aelius, nesse caso, decorre do processador de nomes próprios (Figura 3).

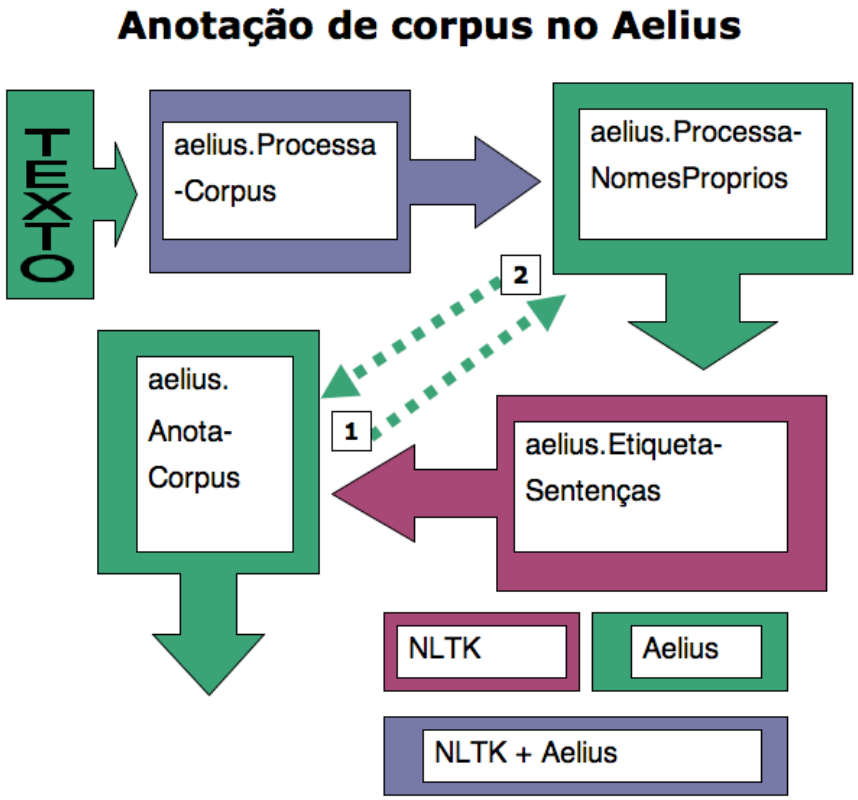

Figura 3: Fluxograma da anotação de corpus no Aelius

Dadas as diferenças profundas entre os tagsets do CHPTB e do corpus Mac-Morpho, base para os etiquetadores on-line do projeto Lácio-Web, uma comparação entre esses e o etiquetador do Aelius é uma tarefa não trivial (Figura 5). Como primeiro passo na comparação entre o Aelius e o TreeTagger do projeto Lácio-Web, extraímos, incialmente, dez sentenças de *Luzia-Homem* aleatoriamente e contamos os erros cometidos pelos dois etiquetadores. Como evidencia a Figura 4, a acurácia do Aelius superou a do TreeTagger em quase 8%. Ressalte-se que o sistema de anotação do CHPTB estabelece mais distinções que o do Lácio-Web (Figura 5).

A tabela 1 faz uma comparação entre os desempenhos do Aelius e de mais quatro etiquetadores na anotação dos três primeiros parágrafos de *Luzia-Homem* (ver os quatro trechos anotados em Alencar, 2010). Constata-se que o Aelius só não supera o TreeTagger que utiliza os parâmetros de Gamallo (2005). No entanto, é importante ressaltar que esse etiquetador utiliza um tagset extremamente simplificado (e bem menos *informativo*, conforme Voutilainen, 2004), que não leva em conta categoriais flexionais nem faz muitas das distinções do tagset do CHPTB (por exemplo, distinção entre nomes comuns e próprios, entre verbos, particípios e gerúndios etc.). Como mostra a Figura 5, o tagset do Aelius, em *Luzia-Homem*, é quase 6 vezes maior do que o desse etiquetador.

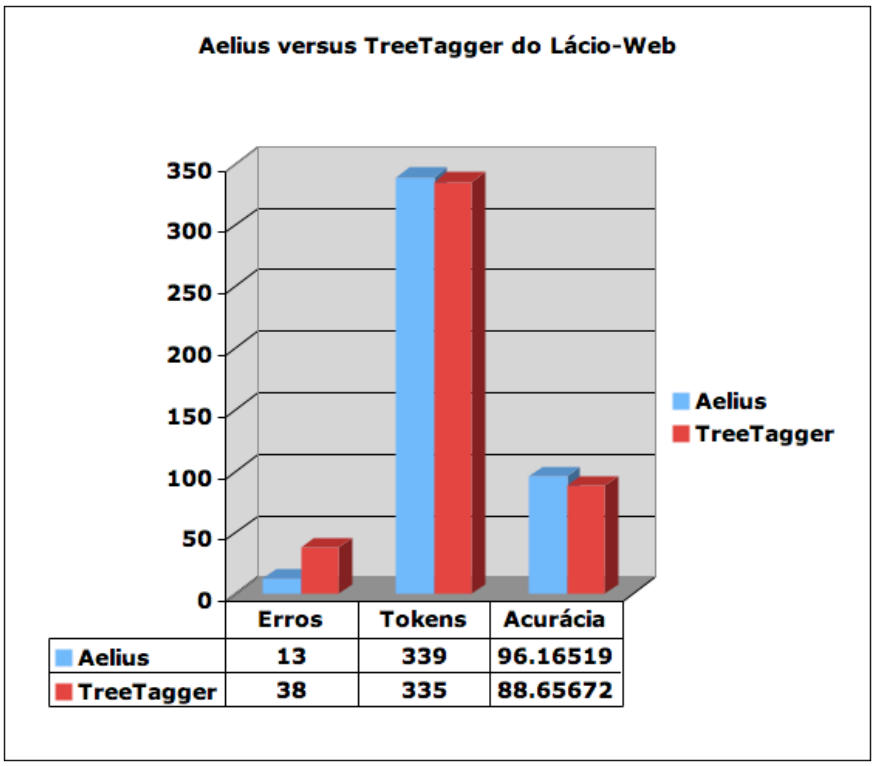

Figura 4: Comparação entre os desempenhos do Aelius e do TreeTagger do Projeto Lácio-Web na anotação de 10 sentenças aleatórias do romance *Luzia-Homem*

Etiquetador	Erros	Tokens	Acurácia
TreeTagger com parâmetros de Gamallo (2005)	3	156	98,08%
Aelius	5	158	96,20%
VLMMTagger treinado em 100% do CHPTB	10	156	93,59%
Brill Lácio-Web	24	165	85,46%
TreeTagger Lácio-Web	26	154	84,15%

Tabela 1: Comparação da acurácia de quatro etiquetadores na anotação dos dois primeiros parágrafos de *Luzia-Homem*.

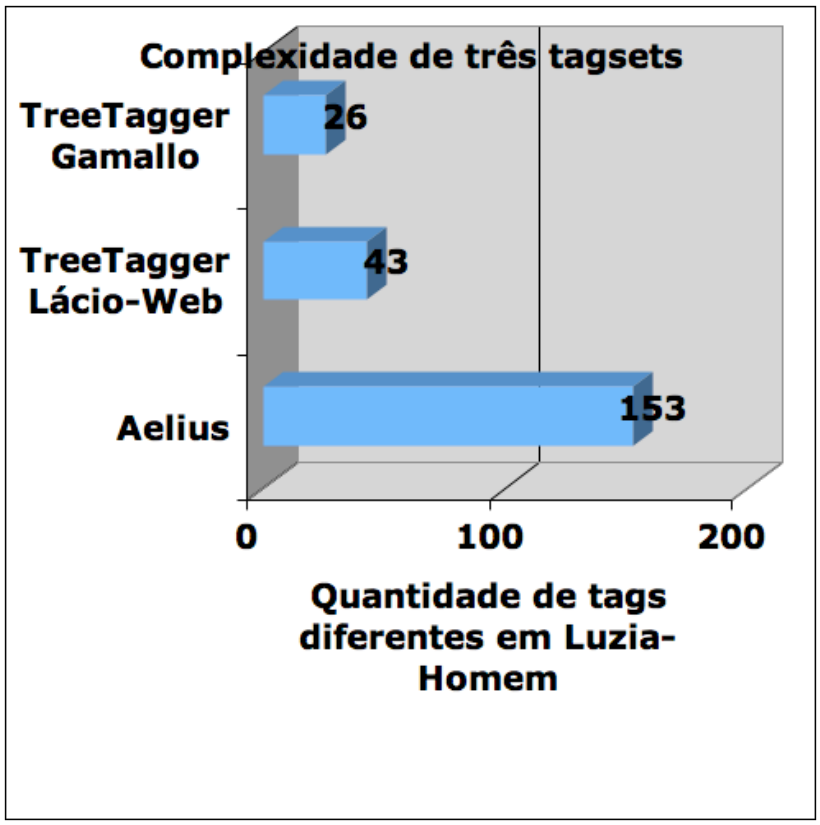

Figura 5: Complexidade dos tagsets dos TreeTaggers de Gamallo (2005), do TreeTagger do Lácio-Web e do Aelius medida pela quantidade de tags na anotação de *Luzia-Homem*

CONCLUSÕES

A análise dos erros cometidos pelo Aelius na anotação dos quatro primeiros capítulos de *Luzia-Homem* indica vias de melhorá-lo, visando a alcançar melhor desempenho quando aplicado ao restante do romance e a outros textos do século XIX. De um lado, parte substancial dos erros envolve palavras de alta freqüência (como, por ex., *retirante*, etiquetado erroneamente como adjetivo), os quais podem ser minimizados treinando o etiquetador nos textos corrigidos manualmente. Outra possibilidade de aumentar a acurácia do etiquetador é diminuir erros que envolvem contextos sintáticos claramente identificáveis, por meio da inclusão de regras de reetiquetagem elaboradas manualmente e/ou pelo encadeamento de um etiquetador de Brill (estratégia adotada por Domingues, Favero e Medeiros, 2008) ou um etiquetador classificador no topo da seqüência de *backoff*, como propõe Perkins (2010), valendo-se do suporte do NLTK à construção de etiquetadores também desses dois tipos.

Acreditamos que o Aelius, cuja acurácia na etiquetagem morfossintática almejamos elevar a 98%, acima, portanto, do estado da arte segundo Garcia e Gamallo (2010), muito contribuirá para a difusão do NLTK nos países de língua portuguesa, ao permitir ampliar consideravelmente o acervo de dados nessa língua distribuídos com a biblioteca. Com isso, dada a amigabilidade do NLTK, estreitará o fosso que ainda separa, entre nós, as comunidades de lingüistas de corpus e lingüistas computacionais.

Por outro lado, o Aelius, aplicado por outros pesquisadores na anotação de textos de gêneros e épocas diversas, fornecerá novos materiais para estudos diacrônicos, dialetológicos ou literários baseados em corpora.

REFERÊNCIAS

ALENCAR, L. F. de. *CORPTEXLIT – Corpus de Língua Portuguesa de Textos Literários do Século XIX*. Fortaleza: [s.n.], 2010. Disponível em:< http://www.leonel.profusehost.net/corptext.html> Acesso em: 30. set. 2010.

BIRD, S.; KLEIN, E.; LOPER, E. *Natural language processing with Python*: analyzing text with the Natural Language Toolkit. Sebastopol, CA: O'Reilly, 2009.

BIRD, S.; KLEIN, E.; LOPER, E. *Natural Language Toolkit*. [s.l]: [s.n.], 2010. Disponível em: <http://www.nltk.org> Acesso em: 30 sep. 2010.

CORPUS Histórico do Português Tycho Brahe. Campinas: Universidade Estadual de Campinas, 2010. Disponível em:<http://www.tycho.iel.unicamp.br/~tycho/corpus/> Acesso em 30. set. 2010.

DOMINGUES, M. L.; FAVERO, E. L.; MEDEIROS, I. P. de. O desenvolvimento de um etiquetador morfossintático com alta acurácia para o português. In: TAGNIN, S. E. O.; VALE, O. A. (Eds.). *Avanços da Linguística de* Corpus *no Brasil*. São Paulo: Humanitas, 2008. p. 267-286.

GAMALLO, P. *Tree-Tagger para português*. Santiago de Compostela: Universidade de Santiago de Compostela, 2005. Disponível em:<http://gramatica.usc.es/~gamallo/> Acesso em: 21 jul. 2010.

GARCIA, M.; GAMALLO, P. Análise morfossintática para português europeu e galego: problemas, soluções e avaliação. *LinguaMÁTICA*, Braga, v. 2, n. 2, p. 59-67, jun. 2010.

KEPLER, F. N. *Um etiquetador morfo-sintático baseado em cadeias de Markov de tamanho variável*. 2005. 70p. Dissertação (Mestrado) – Programa de Pós-Graduação em Ciência da Computação, Universidade de São Paulo, São Paulo, 2005. Disponível em:<http://www.ime.usp.br/~kepler/msc/kepler2005MSc.pdf> Acesso em: 15 abr. 2010.

LÁCIO-WEB: *Ferramentas*. [s.d.]. São Paulo; São Carlos: Universidade de São Paulo. Disponível em:<http://www.nilc.icmc.usp.br/lacioweb/ferramentas.htm> Acesso em: 30 set. 2010.

NAUMANN, S. XML-basierte Tools zur Entwicklung und Pflege syntaktisch annotierter Korpora. In: MEHLER, A.; LOBIN, H. (Eds.). *Automatische Textanalyse*: Systeme und Methoden zur Annotation und Analyse natürlichsprachlicer Texte. Wiesbaden: VS Verlag für Sozialwissenschaften, 2004. p. 153-166.

PERKINS, J. *Part of Speech Tagging with NLTK*. [s.l.]: [s.n], 2010. Disponível em:< http://streamhacker.com/2008/11/03/> Acesso em: 2 out. 2010.

ULE, T.; HINRICHS, E. Linguistische Annotation. In: LOBIN, H.; LEMNITZER, L. (Eds.). *Texttechnologie*: Perspektiven und Anwendungen. Tübingen: Stauffenburg, 2004. p. 217-243.

VOUTILAINEN, A. Part-of-speech tagging. In: MITKOV, R. (Ed.), *The Oxford handbook of computational linguistics*. Oxford, Oxford University Press, 2004. p. 219-232.