Texto Livre Linguagem e Tecnologia
http://periodicos.letras.ufmg.br/index.php/textolivre
Ano: 2014 – Volume: 7 – Número: 2

ANOTAÇÃO MORFOLÓGICA AUTOMÁTICA DE CORPUS DE LÍNGUA FALADA: DESAFIOS AO AELIUS

MORPOHOLOGICAL POS TAGGING IN ORAL LANGUAGE CORPUS: CHALLENGES FOR AELIUS

Gabriel de Ávila Othero/Universidade Federal do Rio Grande do Sul
Mônica Rigo Ayres/Universidade Federal do Rio Grande do Sul

RESUMO: Apresentamos, neste artigo, nosso trabalho de anotação morfológica automática de trechos de um *corpus* de língua falada – pertencentes ao projeto Varsul –, utilizando um etiquetador automático morfossintático gratuito, o Aelius, em 20 textos, perfazendo um total de 154.530 palavras. Basicamente, apresentamos a ferramenta de anotação automática, o processo de análise morfossintática automática efetuada pelo anotador, o trabalho de revisão manual da etiquetagem automática e as sugestões de melhorias para tratar especificamente de aspectos da oralidade. A partir dos erros do etiquetador, buscamos depreender certos padrões de anotação para superar limitações de desempenho apresentadas pelo programa, propondo algumas sugestões de implementações para que o Aelius etiquete de maneira ainda mais satisfatória um *corpus* de língua falada. Tratamos especialmente dos casos de interjeições, aféreses, onomatopeias e marcadores conversacionais.
PALAVRAS-CHAVE: etiquetagem Automática; etiquetagem morfossintática; Linguística de *Corpus*.

ABSTRACT: In this paper, we present the results of our work on automatic morphological annotation using the free morphosyntatic Aelius tagger. The excerpts analyzed were taken from a corpus of spoken language – part of the VARSUL project. We present 20 texts containing 154,530 words, which were automatically annotated and manually corrected. This paper also describes the Aelius tagger and our effort of manual reviewing the texts, as well as our suggestions for improving the tool, concerning aspects of oral texts tagging. We assessed the performance of the tagger at the morphosyntactic tagging of a spoken language corpus, an unprecedented challenge for the tool. Based on its errors, we try to infer certain patterns of annotation to overcome the limitations presented by the program. We also make suggestions in order to allow Aelius to tag spoken language corpora in a more effective way, specially when dealing with interjections, apheresis, onomatopoeia and conversational markers.
KEYWORDS: tagger; morphosyntactic tagging; Corpus Linguistics.

1 Introdução

Nosso texto pretende contribuir com a melhoria de um programa de anotação morfossintática automática, o etiquetador automático Aelius. Analisamos esse anotador automático, gratuito e robusto, propondo melhorias, a fim de que ele possa dar conta de etiquetar corpora de língua falada, além de corpora de língua escrita, que ele já etiqueta

de maneira eficiente. Nossa intenção última é aprimorá-lo para que ele possa ser utilizado para anotação automática de corpora falado, treinando-o com o corpus coletado pelo Varsul, tendo em vista que os textos anotados automaticamente podem auxiliar várias pesquisas linguísticas, tanto de cunho teórico como aplicado.

O trabalho está estruturado e organizado da maneira que segue: a seção 2 discute a fundamentação teórica do trabalho, discorrendo um pouco sobre Linguística de Corpus, corpus e como ele pode ser analisado, com enfoque à etiquetagem morfossintática automática de textos; a seção 3 apresenta a ferramenta computacional Aelius e o corpus de transcrições de entrevistas do Varsul utilizado na pesquisa; a seção 4 traz a análise da correção manual da anotação automática feita pelo Aelius; a seção 5 discute sugestões de implementações para melhoria do etiquetador Aelius em relação à anotação automática de corpus de língua; finalmente, tecemos algumas considerações finais sobre este trabalho.

2 Linguística e *corpus* e etiquetagem automática

A Linguística de Corpus se preocupa em coletar, compilar e explorar conjuntos de textos para pesquisa linguística de uma determinada língua, ou de comparação entre duas ou mais línguas, ou ainda, de uma variedade linguística – e é importante que essa amostra de língua esteja em um formato que seja processável pelo computador. Essa área tem crescido desde que passou a se buscar alternativas para apoiar o estudo da língua em amostras "autênticas" da língua. Os *corpora* justamente permitem que se pesquise e se baseie em linguagem natural e autêntica, possibilitando, assim, a comprovação de teorias por meio de análise dos dados efetivamente produzidos por falantes da língua. Nas palavras de Berber Sardinha (2000a, p. 2),

> 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.

Como vimos, então, um *corpus* é um conjunto de textos ou dados linguísticos autênticos, coletados e selecionados com uma intenção de pesquisa, de maneira que se parte da motivação da pesquisa para o *corpus* e não o contrário. Um *corpus* deve representar determinada variedade linguística, que servirá como base para análise e pesquisa. Segundo Sanchez e Cantos (1996 *apud* Sardinha 2004a, p. 18), um corpus é:

> Um conjunto de dados linguísticos (pertencentes ao uso oral ou escrito da língua, ou a ambos), sistematizados segundo determinados critérios, suficientemente extensos em amplitude e profundidade, de maneira que sejam representativos da totalidade do uso linguístico ou de algum de seus âmbitos, dispostos de tal modo que possam ser processados por computador, com a finalidade de propiciar resultados vários e úteis para a descrição e análise.

Para que um *corpus* seja estudado, é preciso que o pesquisador primeiramente delimite seu enfoque de estudo e só então defina qual é o *corpus* mais adequado para seu objetivo de pesquisa. O *corpus*, independentemente de sua dimensão, deve ser

representativo de uma língua ou de uma variedade dela. Muitas das informações que o linguista quer buscar no *corpus* não estão explícitas; por exemplo, um *corpus* não reconhece o que é "verbo" ou o que é "adjetivo". Esses subsídios precisam ser fornecidos, para que as perguntas que um linguista venha fazer possam ser respondidas, como, por exemplo, quais os adjetivos usados no *corpus*? etc. O fornecimento dessa informação é o que chamamos de **anotação**.

Além da necessidade de escolher bem o *corpus* para sua pesquisa, o pesquisador precisa escolher uma ferramenta que lhe possa ser útil para analisar o *corpus*, dependendo de seu objetivo. Vale destacar a importância do computador para tarefas que poderiam ser muito exaustivas e levar um longo tempo se fossem feitas com nossas limitações humanas, tais como contar palavras. Existem muitas ferramentas que têm capacidade para fazer muitas operações; o que falta, talvez, é conhecimento por parte dos linguistas para utilizar essas ferramentas. Muitos nem as conhecem, muitos não se interessaram por buscar saber usá-las efetivamente. Neste artigo, apresentaremos uma ferramenta que julgamos bastante útil para quem precisa recolher informações morfológicas em um *corpus*: um **etiquetador** morfológico (ou morfossintático).

A **etiquetagem** de um *corpus* pode ser, basicamente, morfológica ou morfossintática (*part of speech tagging*), em que se classificam as unidades lexicais do texto através de *tags* (etiquetas) morfossintáticas; ou sintática (*parsing*), em que se identificam no texto as estruturas sintáticas presentes[1]. Aqui, em nosso texto, trataremos especificamente da etiquetagem morfossintática automática, que marca após cada palavra do *corpus* qual é sua classe gramatical, baseando-se num conjunto de etiquetas morfossintáticas pré-estabelecido.

A etiquetagem de *corpus* é uma parte muito importante nas pesquisas com *corpora*, pois é ela que permite que o *corpus* cumpra seu papel como instrumento para investigação. A etiquetagem morfossintática permite que uma ferramenta computacional etiquete uma grande quantidade de texto de maneira eficiente. De acordo com Alencar (2012, p. 125), a etiquetagem "é uma tarefa aparentemente simples para o processamento da linguagem, no entanto, o desempenho de outras ferramentas depende diretamente desse processo". Além disso, a importância do anotador automático também diz respeito à questão de se entender a linguagem natural, "a etiquetagem morfossintática é uma tarefa intermediária que tem como objetivo principal analisar e entender a língua natural" (ALENCAR, 2012, p. 125).

Para que um texto seja etiquetado, é necessário primeiramente fazer sua toquenização, ou seja, é preciso dividir o texto em *tokens*, que podem ser sentenças ou palavras.O texto normalmente também é "limpo", num processo em que se retiram todas as marcas "não linguísticas" que podem atrapalhar a etiquetagem, por exemplo, apóstrofes, aspas e travessões. Essa limpeza ainda pode ser feita de maneira automática, utilizando-se algum tipo de *script*.

Em síntese, sobre o texto cru, original, é feita a toquenização e a limpeza, ficando o texto pronto para passar pela fase de etiquetagem, em que cada *token* recebe uma etiqueta conforme sua classe gramatical – sempre de acordo com as etiquetas morfossintáticas disponíveis pelo etiquetador.

1 Há também outros tipos de etiquetagem, como a semântica e a discursiva, por exemplo. Sobre as diversas aplicações da Linguística de Corpus, remetemos o leitor a Garside, Leech & McEnery (1997), Berber Sardinha, 2000a, 2000b, 2004a, 2004b, entre outros.

3 O Aelius e o *corpus* anotado

3.1 O Aelius

O projeto do etiquetador Aelius[2] surgiu da necessidade de tornar acessível a estudantes e pesquisadores de Linguística a análise automática de textos, além de ser uma maneira de iniciar estudos na área da Linguística Computacional, por parte de alunos de Engenharia, Computação e Letras. O etiquetador Aelius foi desenvolvido pelo professor Leonel Alencar, da Universidade Federal do Ceará, que coordena o grupo CompLin – Computação e Linguagem Natural. O Aelius pode ser "baixado" gratuitamente no site <http://aelius.sourceforge.net/manual.html> e apresenta uma interface relativamente amigável, como podemos notar abaixo:

```
>>> from Aelius import AnotaCorpus
>>> from Aelius import ProcessaNomesProprios
>>> sents=ProcessaNomesProprios.SENTENCAS
>>> tokens=[sent.split() for sent in sents]
>>> codificadas=AnotaCorpus.codifica_sentencas(tokens)
>>> for s in rubt.batch_tag(codificadas[:3]):
    for w,t in s:
        print "%s/%s " % (w,t),
    print


-/(  Luzia/NPR  pediu/VB-D  a/P  Deus/NPR  e/CONJ  a/D-F  Ávila/NPR  para/P  que/WPRO  lhe/CL  ajudassem/VB-SD  a/D-F  sair/VB  de/P  Sobral
Deus/NPR  ajudou/VB-D  Luzia/NPR  ./.
.../.  Sobral/NPR  era/SR-D  uma/D-UM-F  cidade/N  intelectual/ADJ-G  ./.
.../.  Cidade/NPR  intelectual/ADJ-G  ,/,  Sobral/NPR  tinha/TR-D  muitos/Q-P  poetas/N-P  ./.
Município/NPR  intelectual/ADJ-G  ,/,  Sobral/NPR  tinha/TR-D  muitos/Q-P  poetas/N-P  ./.
Fortaleza/N  era/SR-D  uma/D-UM-F  cidade/N  provinciana/ADJ-F  ./.
Ávila/NPR  ajudou/VB-D  Luzia/NPR  ./.
.../.  Cansada/VB-AN-F  ,/,  Luzia/NPR  logo/ADV  dormiu/VB-D  ./.
Ávida/N  por/P  sossego/N  ,/,  Luzia/NPR  deixou/VB-D  a/P  cidade/N  ./.
Ótimo/N  !/.
Bom/ADJ  ./.
.../.  -/(  Bom/ADJ  ./.
```

Figura 1: Interface do etiquetador automático Aelius.

O Aelius é uma ferramenta desenhada para etiquetar morfologicamente textos escritos de maneira automática, para isso utilizando o mesmo conjunto de etiquetas adotado pelo projeto do *corpus* de português histórico Tycho-Brahe, um *corpus* eletrônico já anotado, composto de textos em português escritos por autores nascidos entre 1380 e 1845. Atualmente, 57 textos (2.547.503 palavras) estão disponíveis para pesquisa livre, com um sistema de anotação linguística em duas etapas: **anotação morfológica** (aplicada em 33 textos) e **anotação sintática** (aplicada em 16 textos) (ALENCAR, 2012, p. 128). Esse *corpus* é desenvolvido junto ao projeto temático Padrões Rítmicos, Fixação de Parâmetros & Mudança Linguística da Universidade Estadual de Campinas[3]. Segundo Alencar (2010, p. 2), "o nível alcançado de acurácia da etiquetagem supera o de ferramentas análogas livremente disponíveis, voltadas sobretudo para o português contemporâneo", ou seja, dentre as ferramentas que temos disponíveis, o Aelius é uma ótima alternativa para a etiquetagem de *corpora*, pois tem um alto índice de acerto, como veremos nas próximas seções.

3.2 Varsul

O projeto Variação Linguística na Região Sul do Brasil – Varsul – estuda,

2 O nome desse etiquetador é uma homenagem ao gramático latino Aelius Donatus, talvez o gramático mais influente de seu tempo. Sobre Aelius Donatus, cf. Holtz (1981) e Dezotti (2010).

3 <http://www.tycho.iel.unicamp.br/~tycho/prfpml/fase2/>. Acesso em: 16 nov. 2014.

justamente, a variação linguística na região Sul do Brasil e conta com a parceria de quatro universidades brasileiras: Universidade Federal do Rio Grande do Sul, Pontifícia Universidade Católica do Rio Grande do Sul, Universidade Federal de Santa Catarina e Universidade Federal do Paraná. Esse projeto foi criado no ano de 1982, seguindo a proposta da professora Leda Bisol, e tem como objetivos principais proporcionar subsídios para descrever a língua falada no Brasil; criar condições para teste e desenvolvimento de teorias linguísticas; e oferecer subsídios para programas educacionais, promovendo, assim, o conhecimento e o respeito às variedades linguísticas[4]. Atualmente, o projeto do Varsul se divide em três: Banco de Dados do VARSUL, Amostra Digital VARSUL e Banco de Dados Diacrônico[5].

De acordo com Raso e Mello (2012), há três tipologias de interação em entrevistas em fala espontânea: a monológica, que ocorre quando um falante sozinho fala sem interação; a dialógica, em que dois falantes conversam entre si; e a tipológica conversacional, que ocorre quando dois ou mais falantes interagem. Todas as entrevistas do Varsul que analisamos para este trabalho foram de interação dialógica, registrando um diálogo entre um aluno da UFRGS, bolsista do Varsul, e a pessoa entrevistada.

Efetuamos uma análise morfológica, automática e manualmente, como veremos, de vinte trechos de entrevistas transcritas do Banco de Dados do Varsul, todas ocorridas em Porto Alegre, entre os anos de 1990 e 1999. Esses vinte trechos totalizam 410 páginas e 154.530 palavras. Na tabela abaixo, temos esquematizado nosso *corpus*, apresentando o nome de referência de cada trecho do Varsul, além de seu número de páginas e de palavras.

Tabela 1: Constituição do *corpus*.

REFERÊNCIA VARSUL	Nº de PÁGINAS	Nº de PALAVRAS
POA 01	7	3.338
POA 01.1	23	9.572
POA 02	8	3.285
POA 02.1	31	11.572
POA 03	8	2.965
POA 03.1	27	9.919
POA 04	17	6.877
POA 05	7	2.923
POA 12	8	2.797
POA 15	8	2.691
POA 21	9	3.338
POA 25	29	10.890
POA 26	28	10.152
POA 27	30	10.426
POA 28	29	10.510
POA 29	30	11.908
POA 31	29	11.467
POA 32	30	11.880

4 <http://www.varsul.org.br/>. Acesso em: 16 nov. 2014.

5 As entrevistas do Varsul estão disponíveis nas universidades participantes do projeto.

POA 35	**18**	**7.397**
POA 41	**34**	**10.623**
TOTAL	**410**	**154.530**

A partir desse *corpus* e com a ferramenta apresentada, o Aelius, partimos, na próxima seção, para as análises da etiquetagem automática morfossintática e para a esquematização de alguns dos principais problemas encontrados no processo de anotação.

4 A anotação

4.1 Processamento dos textos

Primeiramente, foram selecionados alguns trechos de transcrições de entrevistas do banco de dados de Porto Alegre do Varsul, aleatoriamente. Selecionados os textos, que estavam em formato .doc, foi necessário passá-los para o formato .txt, tendo em vista que é o formato que o etiquetador Aelius espera receber seus arquivos para poder rodar os textos e fazer a etiquetagem.

Esses trechos passaram por uma limpeza que retirou todas as marcas que poderiam atrapalhar a anotação, como travessões, asteriscos e parênteses, por exemplo. Após os textos estarem “limpos”, eles foram, então, rodados no etiquetador automático Aelius. Finalmente, foram impressos e revisados manualmente.

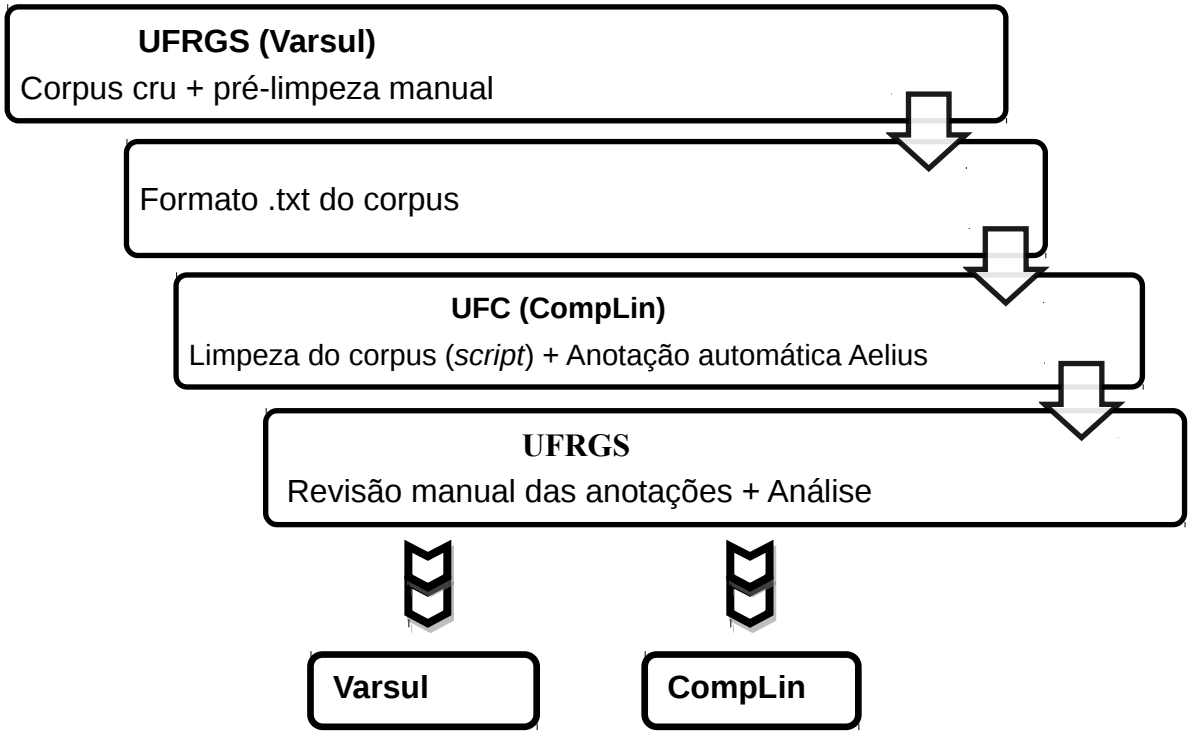

Figura 2: Processo de anotação dos textos.

4.2 Exemplos de textos

Neste primeiro quadro, trazemos um trecho como exemplo de texto cru, em que notamos a presença de caracteres que podem atrapalhar o etiquetador, como asteriscos, sustenidos e colchetes.

... entregar pra amanhã, vai se entregar, se Deus quiser, salvo algum problema, né? extra, ma0s está tudo bem, vai ser entregue, se Deus quiser. *Então, [isso]-[isso]- sabe que isso é muito importante, sabe o que que é? *Então a pessoa fa0z uma previsão: "*Eu vou entregar o serviço-" *Pra ti no caso: "*Olha, eu vou entregar teu carro sexta#feira." *Se eu não entregar sexta#feira vai ficar meio ridículo, meio xarope, sabe? *Você vai entender, ma0s no fundo, no fundo você: "*Pô ma0s Eduardo0 não cumpriu com a palavra." *No seu íntimo, né?...

Quadro 1: Exemplo de texto original (Varsul).

O segundo quadro é uma amostra de como o texto fica após ser “limpo” pelo *script*, sem as marcas que podem interferir na etiquetagem.

... entregar pra amanhã, vai se entregar, se Deus quiser, salvo algum problema, né? extra, mas está tudo bem, vai ser entregue, se Deus quiser. Então, isso isso sabe que isso é muito importante, sabe o que que é? Então a pessoa faz uma previsão: "Eu vou entregar o serviço" Pra ti no caso: "Olha, eu vou entregar teu carro sextafeira." Se eu não entregar sextafeira vai ficar meio ridículo, meio xarope, sabe? Você vai entender, mas no fundo, no fundo você: "Pô mas Eduardo não cumpriu com a palavra." No seu íntimo, né?...

Quadro 2: Exemplo de texto “limpo”.

No terceiro quadro, temos um exemplo de como o texto fica após ser etiquetado automaticamente pelo Aelius. Ao ser anotada, cada palavra recebe um símbolo / (barra) e após esse símbolo uma etiqueta que define sua classe, como podemos observar abaixo.

... entregar/VB pra/P amanhã/ADV ,/, vai/VB-P se/SE entregar/VB-SR ,/, se/CONJS Deus/NPR quiser/VB-SR ,/, salvo/P algum/Q problema/N ,/, né/NPR ?/. Extra/ADV ,/, mas/CONJ está/ET-P tudo/Q bem/ADV ,/, vai/VB-P ser/SR entregue/VB-AN ,/, se/CONJS Deus/NPR quiser/VB-SR ./. Então/ADV ,/, isso/DEM isso/DEM sabe/VB-P que/C isso/DEM é/SR-P muito/Q importante/ADJ-G ,/,

sabe/VB-P o/D que/WPRO que/WPRO é/SR-P ?/. Então/ADV a/D-F pessoa/N faz/VB-P uma/D-UM-F previsão/N :/. "/QT Eu/PRO vou/VB-P entregar/VB o/D serviço/N "/QT Pra/P ti/PRO no/P+D caso/N :/. "/QT Olha/VB-P ,/, eu/PRO vou/VB-P entregar/VB teu/PRO$ carro/N sextafeira/NPR ./. "/QT Se/CONJS eu/PRO não/NEG entregar/VB sextafeira/N vai/VB-P ficar/VB meio/N ridículo/ADJ ,/, meio/ADJ xarope/N ,/, sabe/VB-P ?/. Você/PRO vai/VB-P entender/VB ,/, mas/CONJ no/P+D fundo/N ,/, no/P+D fundo/N você/PRO :/. "/QT Pô/NPR mas/CONJ Eduardo/NPR não/NEG cumpriu/VB-D com/P a/D-F palavra/N ./. "/QT No/P+D seu/PRO$ íntimo/N ,/, né/NPR ?/. ...

Quadro 3: Exemplo de texto anotado automaticamente pelo Aelius[6].

A correção manual da etiquetagem automática é um trabalho que exige muita atenção e revisão, pois são muitas as etiquetas que podem aparecer, e uma mesma palavra pode aparecer com etiquetas diferentes, dependendo dos contextos em que estiver inserida. Por isso, deve-se prestar atenção ao contexto da palavra e não apenas à etiqueta da palavra isolada ou pensada em um contexto abstrato.

Neste quadro, podemos ter uma ideia de como se dá a revisão manual da etiquetagem automática, pois aqui trazemos um exemplo de texto que foi devidamente "limpo", etiquetado automaticamente e revisado manualmente. As etiquetas marcadas equivocadamente estão destacadas em negrito.

... entregar/VB pra/P amanhã/ADV ,/, vai/VB-P se/SE ***entregar*/VB-SR** ,/, se/CONJS Deus/NPR quiser/VB-SR ,/, ***salvo*/P** algum/Q problema/N ,/, ***né*/NPR**?/. ***Extra*/ADV** ,/, mas/CONJ está/ET-P tudo/Q bem/ADV ,/, vai/VB-P ser/SR entregue/VB-AN ,/, se/CONJS Deus/NPR quiser/VB-SR ./. Então/ADV ,/, isso/DEM isso/DEM sabe/VB-P que/C isso/DEM é/SR-P muito/Q importante/ADJ-G ,/, sabe/VB-P o/D que/WPRO que/WPRO é/SR-P ?/. Então/ADV a/D-F pessoa/N faz/VB-P uma/D-UM-F previsão/N :/. "/QT Eu/PRO vou/VB-P entregar/VB o/D serviço/N "/QT Pra/P ti/PRO no/P+D caso/N :/. "/QT Olha/VB-P ,/, eu/PRO vou/VB-P entregar/VB teu/PRO$ carro/N ***sextafeira*/NPR** ./. "/QT Se/CONJS eu/PRO

6 As etiquetas que aparecem neste trecho são as seguintes: ADJ adjetivo; ADJ-G adjetivo neutro; ADV advérbio; C complemetizador; CONJ conjunção; CONJS conjunção; D demonstrativo; D-F demonstrativo feminino; DEM demonstrativo; D-UM-F demonstrativo um feminino; N nome; NEG negação; NPR nome próprio; SR verbo ser infinitivo; P preposição; P+D pronome + demonstrativo; PRO pronome; POR$ pronome possessivo; Q quantificador; QT aspas; SE se; SR-P verbo ser presente; VB verbo infinitivo; VB-P verbo presente; VB-SR verbo futuro subjuntivo; e, WPRO pronome relativo ou interrogativo. A lista de etiquetas original pode ser visualizada no endereço <http://www.tycho.iel.unicamp.br/~tycho/corpus/manual/tags.html>. Acesso em: 16 nov. 2014.

não/NEG entregar/VB sextafeira/N vai/VB-P ficar/VB **meio/N** ridículo/ADJ ,/, meio/ADJ ***xarope*/N** ,/, sabe/VB-P ?/. Você/PRO vai/VB-P entender/VB ,/, mas/CONJ no/P+D fundo/N ,/, no/P+D fundo/N você/PRO :/. "/QT ***Pô*/NPR** mas/CONJ Eduardo/NPR não/NEG cumpriu/VB-D com/P a/D-F palavra/N ./. "/QT No/P+D seu/PRO$ íntimo/N ,/, ***né*/NPR**?/. ...

Quadro 4: Exemplo de texto anotado automaticamente pelo Aelius e revisado manualmente.

Para fazer a correção manual desse *corpus*, analisamos os 20 textos destacando as etiquetas que apareceram erradas. Após essa primeira análise, revisamos os textos para garantir que nenhuma etiqueta passasse despercebida, principalmente se empregada incorretamente.

A Tabela 2, a seguir, ilustra a composição do *corpus* e a acurácia das análises do Aelius. Na primeira coluna, "Referência Varsul", temos o nome do arquivo que nos foi fornecido do banco de dados do Varsul, das transcrições das entrevistas. Nas linhas em que temos o mesmo número de referência, foram analisados dois trechos da mesma entrevista. Na segunda coluna, temos a informação do número de palavras totais de cada trecho e, na coluna seguinte, o número de palavras que foram etiquetadas equivocadamente. Na terceira coluna, calculamos a acurácia do anotador em cada um dos trechos individualmente. Na última linha, temos a totalização dos números apresentados na tabela, que são os números totais do *corpus* analisado, contendo 154.530 palavras, das quais 7.004 foram etiquetadas equivocadamente, de maneira que a acurácia do Aelius é de 95,4% em nosso *corpus*.

Tabela 2: Acurácia do Aelius na nossa amostra do *corpus*.

REFERÊNCIA VARSUL	Nº de PALAVRAS	Nº de ERROS	ACURÁCIA
POA 01	3.338	203	93,9%
POA 01.1	9.572	508	94,6%
POA 02	3.285	135	95,8%
POA 02.1	11.572	398	96,5%
POA 03	2.965	96	96,7%
POA 03.1	9.919	351	96,4%
POA 04	6.877	252	96,3%
POA 05	2.923	108	96,3%
POA 12	2.797	121	95,6%
POA 15	2.691	189	92,9%
POA 21	3.338	180	94,6%
POA 25	10.890	466	95,7%
POA 26	10.152	419	95,8%
POA 27	10.426	337	96,7%
POA 28	10.510	456	95,6%
POA 29	11.908	532	95,5%
POA 31	11.467	824	92,8%

POA 32	**11.880**	**432**	**96,3%**
POA 35	**7.397**	**531**	**92,8%**
POA 41	**10.623**	**466**	**95,6%**
TOTAL	**154.530**	**7.004**	**95,4%**

O Aelius tem uma acurácia considerada alta (95,4%). Entretanto, sua eficiência é ainda maior na análise de textos escritos: 96,3% de acurácia (cf. ALENCAR, 2013). Para efeitos de comparação, o etiquetador LX-Tagger[7] (uma ferramenta desenvolvida na Universidade de Lisboa pelo NLX - Grupo de Fala e Linguagem Natural do Departamento de Informática) tem 96,2% de acurácia em textos de língua escrita. O LX-Tagger é um etiquetador morfossintático como o Aelius e etiqueta seus *corpora* de forma semelhante, utilizando o símbolo / (barra) após cada ocorrência e definindo sua devida etiqueta.

4.3 Alguns problemas encontrados na correção manual

O Aelius foi desenhado para analisar textos de língua escrita; por isso, encontramos alguns problemas ao utilizá-lo para etiquetar textos de língua falada, como esperado. Na língua falada temos ruídos, hesitações e truncamentos que acabam gerando palavras desconhecidas pelo etiquetador, que precisa encontrar uma solução em seu dicionário para gerar uma etiqueta para essas "pseudo-palavras"; há algumas interjeições menos comuns e expressões que o anotador tem dificuldade de etiquetar; e, além disso, há gírias e palavras típicas e específicas (de Porto Alegre, no nosso caso), que acabaram sendo etiquetadas equivocadamente. Outros problemas encontrados, que não se restringem a limitações do programa frente a dados de língua falada, dizem respeito à estrutura de nome próprio composto (que o anotador simplesmente não reconhece) e à confusão entre substantivos e adjetivos. Aqui, neste artigo, pretendemos apresentar uma sistematização dos erros do Aelius oriundos de características da língua falada. Lembramos que o Aelius foi treinado em um corpus de língua escrita de português histórico – *corpus* do projeto Tycho Brahe. É mais do que natural, portanto, que o etiquetador "sofra" para analisar um *corpus* transcrito de língua falada contemporânea. De qualquer maneira, como já ressaltamos, ainda assim o percentual de acerto do etiquetador foi bastante elevado (95,4%) ao analisar nosso *corpus* de língua falada. Nossa sistematização de erros pretende servir de base para melhorias no Aelius, a fim de que ele possa ter um índice ainda mais apurado na análise de outros *corpora* de língua falada.

4.4 Erros de anotação da fala

Por o Aelius ter sido projetado para ser (e treinado como) um etiquetador para anotação automática de *corpora* em língua escrita, ele encontrou alguma dificuldade em analisar expressões típicas da fala, como marcadores conversacionais, hesitações e truncamentos. Abaixo, nas tabelas que seguem, destacamos os erros desse tipo que

7 <http://lxcenter.di.fc.ul.pt/tools/pt/conteudo/LXTagger.html#pq>. Acesso em: 16 nov. 2014.

foram encontrados durante a correção manual dos trechos anotados pelo etiquetador e apresentamos o contexto em que esses erros ocorreram. Na próxima seção, propomos etiquetas que sejam adequadas para descrever essas expressões.

Tabela 3: Alguns erros de anotação da fala.

OCORRÊNCIA	CONTEXTO
Ãh	... como/CONJS é/SR-P que/C se/SE diz/VB-P ,/, ***ãh*/N** Ah/INTJ !/.
	... morar/VB num/P num/P centro/N mais/ADV-R ***ãh*/NPR** assim/ADV
Ahã	... isto/DEM ***ahã*/ADV** ,/, são/SR-P daqui/P+ADV
Bé	... ovelhinha/N fazia/VB-D “/QT ***bé*/NPR** “/QT ,/, “/QT ***bé*/NPR** “/QT
Bom	... eu/PRO digo/VB-P :/. ***Bom*/ADJ** ,/. vou/VB-P ter/TR-P que/WPRO
Claro	... muito/Q gaúcho/N ,/, ***claro*/ADJ** ,/, *né*/NPR ?/.
Hum	... próximas/ADJ-F-P eleições/N-P ?/. ***Hum*/D-UM** ./. E/CONJ o/D
Hum-hum	... aqui/ADV do/P+D lado/N ?/. ***Hum*/D-UM *hum*/D-UM** ./. Incluisve
Né	... era/SR-D hoje/ADV o/D sem/P fio/N ,/, ***né*/NPR** ?/.
Olha	... F/NPR ***Olha*/VB-P** ,/, *chimarrão*/VB-R a/D-F gente/N fazia
Tá	... aí/ADV ***tá*/ADV** ,/, mas/CONJ a/D-F guria/N nervosa/ADJ-F
	... o/D cara/N ,/, ***tá*/NPR** ,/, tudo/Q o/D cara/N
Tal	... todo/Q mundo/N no/P+D banho/N e/CONJ ***tal*/ADJ-R-G**
Tarará	... trabalho/N ,/, (/(“/QT ***tarará*/NPR** “/QT)/(,/, leva/VB-P trabalho/N
Tipo	... coisas/N-P lá/ADV assim/ADV ***tipo*/N** ,/, quando/CONJS eu/PRO

Por se tratar de um *corpus* de língua falada, o Aelius enfrenta dificuldade para reconhecer truncamentos, que acabam gerando etiquetas diversas, pois não há a etiqueta “truncamento”. Para mencionar alguns exemplos, temos “caind”, “cans”, “di”, “el”, “mu”, “mour”, “maq”, “maqui”, “su” e sub. Essas “palavras” receberam as mais variadas etiquetas: nome próprio, nome plural, verbo, advérbio etc. Na tabela 4, mostraremos os contextos em que ocorreram:

Tabela 4: Exemplos de truncamento.

TRUNCAMENTO	CONTEXTO
caind	... ,/, acaba/VB-P caind/P ***caindo*/VB-G** em/P domínio/N público/ADJ
cans	... tínhamos/TR-D ***cans*/N-P** *calçamento*/NPR e/CONJ não/NEG
di	... tenha/VB-SR opções/N-P ***di*/VB-P** diferentes/ADJ-G-P ,/, quero
el	... compensar/VB o/D fato/N de/P ***el*/NPR** dela/P+PRO não/NEG

maq/maqui **mu/mour** **su/sub**	estar ... ***maquiar*/VB *maqui*/P+ADV** maq/VB-D aí/ADV já/ADV é/SR-P ... nós/PRO fomos/VB-D ***mu*/VB-D *mour*/VB** fomos/VB-D morar ... matérias/N-P foram/SR-D ***su*/VB-D *sub*/N-P** suprimidas/VB-NA-F-

Como podemos observar, *cans* foi etiquetada como nome plural, possivelmente por terminar em *-s*. A palavra seguinte, *calçamento*, foi etiquetada como nome próprio, e isso provavelmente aconteceu porque as etiquetas se relacionam umas com as outras, ou seja, se antes da palavra *calçamento* houvesse um determinante e não um nome plural, a probabilidade de ser marcada corretamente como nome seria maior. Segundo nossa análise, aparentemente os outros exemplos de truncamento não influenciaram em etiquetas erradas em outros vocábulos.

Acreditamos que casos que lidam com truncamentos sejam complicados de se resolver, pois é bastante complicado que o etiquetador reconheça quando se trata de uma palavra ou de um "pedaço" de palavra. Uma maneira possível de resolver o problema com palavras truncadas pode ter alguma relação com o algoritmo elaborado por Alencar (2009). Nesse texto, Alencar propõe um reconhecedor de neologismos. O truncamento obviamente não é um neologismo, mas é uma "palavra", de certa maneira, nova. O anotador deve ser capaz de reconhecer quando uma palavra aparece ao lado de um fragmento que repete sua estrutura, antecedendo-a (como no exemplo *su sua*) ou sucedendo-a (como no exemplo *maquiagem maqui maq*), que destacamos na Tabela 4.

5 Proposta de novas etiquetas

Nesta seção, sugerimos implementações para que o Aelius etiquete com ainda mais eficiência *corpus* de língua falada. O Aelius – ou qualquer outro etiquetador – não vai necessariamente acertar a etiqueta de uma mesma palavra em todas as suas ocorrências, pois qualquer mudança no contexto sintático da palavra a ser etiquetada pode bastar para que o etiquetador decida por outra etiqueta. Temos de levar em conta que a análise morfossintática que nós, linguistas, fazemos considera uma grande quantidade de regras e padrões que nos são naturais, mas, ao etiquetador, foram passadas apenas algumas dessas informações, através de moldes estatísticos; por isso, algumas vezes o etiquetador acaba sendo induzido a etiquetar equivocadamente, pois ele não conta com a intuição de falante natural.

5.1 Sugestões e novas etiquetas

Com base na análise que fizemos da anotação do Aelius, percebemos que seria muito valiosa a inclusão de novas etiquetas que pudessem fazer com que o índice de acertos do anotador fosse ainda mais alto. Para fazermos implementações em um etiquetador, devemos utilizar um *corpus* já anotado e revisado, para que se treine um algoritmo de aprendizagem de máquina. Os algoritmos utilizam-se de regras estatísticas

para aprender como as etiquetas se relacionam entre si. Exemplificando, a frequência de determinantes antes de substantivos é muito grande; por isso, há grandes chances de a palavra que esteja antes de um nome ser um determinante e vice-versa, e o anotador adquire essa noção através do que é "ensinado" a ele, com regras estatísticas e com a inserção de palavras no dicionário do programa. Segundo Voutilainen (2009 *apud* Alencar 2012, p. 126),

> essas regras podem ser baseadas sobre duas fontes de informações, ambas codificadas no etiquetador na forma de uma linguagem modelo: a informação sobre a palavra em si, ou seja, em que contexto efetivo a palavra é mais usada, por exemplo, a palavra como: verbo ou advérbio; e as informações sobre a sequência da palavra palavra/etiqueta (ou contexto informacional): isto é, o modelo pode preferir analisá-lo como um verbo a uma conjunção, se o termo precedente for um advérbio ou um determinante.

Para a inserção de novas etiquetas, é necessário o treinamento de um algoritmo de aprendizagem de máquina com um *corpus* que tenha passado por anotação e revisão com essas novas etiquetas, para que se arquitete um novo etiquetador. Também aqui reside a importância de nosso trabalho: disponibilizamos um corpus de língua falada etiquetado automaticamente e revisado manualmente. Não só o Aelius poderá ser treinado com base nesse corpus anotado, mas também qualquer outro etiquetador morfológico em língua portuguesa que pretenda etiquetar *corpora* falados.

Além disso, vamos aqui sugerir novas etiquetas ao Aelius, baseando-nos em nossa análise linguística dos dados que levantamos a partir da correção manual, sistematizados nas tabelas da seção anterior. Com a implementação dessas etiquetas, esperamos que a acurácia do Aelius seja ainda maior em sua tarefa de anotação de língua falada.

5.2 Inserção de novas etiquetas

Para que o Aelius possa dar conta de todas as palavras do *corpus* analisado – e, futuramente, de outros *corpora* de língua falada –, pensamos que é necessário que haja uma etiqueta que dê conta de casos muito específicos da fala, como as onomatopeias, as aféreses e os marcadores conversacionais.

A onomatopeia é a criação de uma palavra que reproduz o mais "fielmente" possível um som. Em nossa análise, ocorreu apenas um caso de onomatopeia, a imitação do som de uma ovelha: ***bé.*** Para que o Aelius dê conta dessa e de outras onomatopeias que possam aparecer em *corpora* futuros, acreditamos que seja adequado inserir a etiqueta "**ONTP**" - Onomatopeia, para que o anotador etiquete corretamente as palavras como as seguintes: *argh, atchim, au, bang, bé, blá, brr, bzz, clique, cof, grr, hum hum, nhac, tec, tic, tac, toc, plaft, piu, pof, pum* e *vrum*. Com o passar do tempo, se surgir a necessidade, essa lista pode vir a aumentar. Esperamos que assim o Aelius dê conta desse recurso de formação de palavra, não muito produtivo, mas encontrado nas línguas naturais, especialmente na modalidade falada.

Outro caso que apareceu no corpus foram ocorrências de aférese, a supressão de fonema no início das palavras. No *corpus* do Varsul analisado, ocorreram apenas casos de aférese verbal, com os quais o etiquetador conseguiu lidar satisfatoriamente, como ***tá***,

tão etc. No entanto, como pretendemos propor melhorias para que o Aelius possa analisar textos na modalidade falada da língua, de maneira irrestrita, acreditamos que seja válido preparar o anotador para lidar com casos de aférese nominal, adjetival, preposicional e adverbial, como por exemplo: “fessora” (professora), “brigada” (obrigada), té (até), “bora” (embora), etc.

Finalmente, nenhum anotador que pretenda analisar *corpora* falados conseguirá analisar morfologicamente o *corpus* de maneira adequada sem levar em consideração os marcadores conversacionais, que são aquelas palavras ou expressões – típicas e corriqueiras da fala, principalmente da conversa – que funcionam como subsídios para a interação. Há marcadores utilizados pelo ouvinte e pelo falante, e os dois apoiam-se mutuamente para que o diálogo continue. Os marcadores são utilizados para marcar hesitação, servem como busca de apoio do interlocutor e funcionam como monitoramento do ouvinte. Segundo Urbano (2010, p. 93), os marcadores conversacionais são

> elementos de variada natureza, estrutura, dimensão, complexidade semântico-sintática, aparentemente supérfluos ou até complicadores, mas de indiscutível significação e importância para qualquer análise de texto oral e para sua boa e cabal compreensão.

Esses marcadores podem ser linguísticos ou não. Os marcadores não linguísticos (ou “extralinguísticos”) são ações como risos e olhares, por exemplo, e não tratamos deles aqui, pois não temos como captá-los, justamente por não serem marcas linguísticas *per se*[8].

Os marcadores conversacionais linguísticos se dividem em prosódicos e verbais. Os marcadores prosódicos também não serão analisados neste trabalho, porque são os alongamentos de sílabas, as pausas e as mudanças na tessitura ou na velocidade da voz, ou seja, são aqueles aspectos que não estão marcados no *corpus* e só conseguiriam ser percebidos se ouvíssemos as entrevistas. Os marcadores verbais são aspectos que podemos notar na transcrição das entrevistas e se dividem em lexicalizados e não lexicalizados, segundo a classificação proposta por Urbano (2010). Tanto os lexicalizados como os não lexicalizados tiveram ocorrência em nosso *corpus*. Os lexicalizados são *Bom*, *Claro*, *Tipo* etc., e os não lexicalizados são *Ãh*, *Ahã* e *Hum*, por exemplo. Urbano ainda fala sobre a diferença dos marcadores em relação a seu lugar na sentença, dizendo que alguns têm mais tendência a ocorrer no início, como *Bom* e *Olha*; outros têm a tendência de ocorrer no meio da frase, como *E tal* e *Claro*; e outros, ainda, têm maior probabilidade de ocorrer ao final, como *Tá* e *Né*. Em nosso *corpus* tivemos ocorrências de todos esses tipos e vamos analisá-las a seguir.

- Marcadores conversacionais não lexicais:

Ãh, **Ahã**, **Hum**, **Hum hum** e **Tarará** foram os casos que ocorreram em nosso *corpus*; por isso, sugerimos que eles sejam adicionados ao dicionário do etiquetador com a etiqueta **MCnL** de marcador conversacional não lexical. Além dessas ocorrências, sugerimos que sejam adicionadas palavras como **Eh**, **Ahn** e **Uhn**, que também podem aparecer em *corpora* futuros.

8 Na verdade, há registro de risos no corpus analisado, mas escrito como “risos”, e etiquetado como nome plural.

- Marcadores conversacionais lexicais:

Bom, **Claro**, **Né**, **Olha**, **Tá**, **Tal** e **Tipo** são ocorrências do nosso *corpus*; portanto, sugerimos que essas palavras sejam adicionadas com a etiqueta **MCL** de marcador conversacional lexical. Aqui há de se tomar cuidado, pois como todas são palavras lexicais, elas têm mais possibilidades de significação. Por isso, o algoritmo precisa ser bem treinado, levando em conta as que aparecem no início da frase, as mediais e as finais (*Olha*, por exemplo, tem alta probabilidade de ser um MCL quando for a primeira palavra de uma frase).

Em comunicação pessoal, Sergio Menuzzi (UFRGS) sugeriu que talvez os marcadores conversacionais que aparecem em posição final podem estar sendo confundidos com nomes próprios – como efetivamente verificamos nos trechos em que o Aelius anotou certos marcadores conversacionais equivocadamente nessa posição – porque essa é uma posição típica de vocativo, ocorrendo sempre ao final da sentença e geralmente após uma vírgula, seguida por um ponto de exclamação ou interrogação. Por exemplo: o trecho que encontramos ***fazer rancho e tudo mais, né?*** poderia facilmente ser substituído por um trecho com um nome próprio: “fazer rancho e tudo mais, João?”, ainda que a força ilocutiva da sentença tenha sido alterada – algo que passa despercebido pelo Aelius. Tanto isso parece ser assim que a análise feita pelo Aelius desse trecho foi a seguinte: *fazer/VB rancho/N e/CONJ tudo/Q mais/ADV-R ,/,* ***né/NPR*** *?/.*, onde a etiqueta NPR significa “nome próprio”. Esse raciocínio deve ser levado em conta quando a posição dos marcadores for estudada.

6 Considerações finais

A Linguística de Corpus vem evoluindo cada vez mais ao longo dos últimos anos, mas um problema frequentemente encontrado é que muitas vezes não há ferramentas disponibilizadas gratuitamente para que os pesquisadores possam ter mais eficiência e versatilidade em seus objetivos de pesquisa. Com a finalidade de auxiliar pesquisadores em linguística, o Aelius foi criado, e a fim de aumentarmos seu escopo de anotação automática morfossintática para além de língua escrita, abrangendo também língua falada, fizemos este trabalho, analisando exaustiva e cuidadosamente o *corpus* de língua falada anotado automaticamente.

Esperamos que nossa pesquisa possa contribuir de maneira satisfatória com a melhoria desse programa que pode auxiliar muitas pesquisas em nossa área. Além disso, esperamos que cada vez mais os profissionais da linguagem tenham conhecimento dessas ferramentas que estão a nossa disposição e que efetivamente as usem, pois são instrumentos valiosos para a pesquisa. A importância da Linguística de Corpus, em conjunto com a Linguística Computacional, pode ser constatada na “elaboração de teorias gramaticais formalmente mais consistentes e psicolinguisticamente mais realistas [...] e, assim, testar, com um grau de sofisticação que dificilmente poderia ser atingido por seres humanos, a adequação dos modelos postulados” (ALENCAR; OTHERO, 2011, p. 9).

Em nosso trabalho, iniciamos falando sobre o que é a Linguística de Corpus e o

que é um *corpus*, dando exemplos de como um *corpus* pode ser analisado, focando principalmente no assunto de nossa pesquisa, a anotação automática morfossintática.

Depois, apresentamos a ferramenta utilizada, o Aelius, e o nosso *corpus*, constituído por trechos de transcrição de entrevistas do Banco de Dados do Varsul. Após apresentarmos isso, passamos para a parte mais exaustiva do trabalho, a análise dos dados obtidos com a correção manual da anotação automática feita pelo anotador Aelius, na qual destacamos os principais problemas encontrados na correção e separamos as ocorrências por tipos de erros (confusão de etiquetas ou erros de anotação da fala), mostrando os contextos em que ocorreram.

Com a intenção de adequarmos o Aelius para outros *corpora* de língua falada, além de melhorar a anotação do *corpus* do Varsul, sugerimos algumas implementações de etiquetas (onomatopeias, aférese e marcadores conversacionais) e adição de novas palavras ao dicionário do etiquetador, no grupo das etiquetas das interjeições.

Com isso, esperamos que a ferramenta seja aperfeiçoada em uma próxima versão e, além disso, que seja mais utilizada por linguistas, pois esperamos que o Aelius obtenha maior número de acertos – não esquecendo que ele já tem uma acurácia alta – e que o Varsul tenha à sua disposição um etiquetador automático morfossintático de qualidade, para futuras pesquisas.

Acreditamos que essas ferramentas computacionais que auxiliam o estudo da língua poderiam ser ainda mais usadas por linguistas porque há muitos *corpora* disponíveis e as ferramentas estão cada vez mais acessíveis, tornando nossas pesquisas mais eficientes e confiáveis.

Referências

ALENCAR, Leonel Figueiredo de. Produtividade morfológica e tecnologia do texto: aspectos da construção de um transdutor lexical do português capaz de analisar neologismos. *Calidoscópio* (UNISINOS), v. 7, 2009.

ALENCAR, Leonel Figueiredo de. Aelius: uma ferramenta para anotação automática de corpora usando NLTK. *IX Encontro de Linguística de Corpus***.** Porto Alegre, PUCRS, 2010.

ALENCAR, Leonel Figueiredo de. Superando o estado da arte na etiquetagem morfossintática por meio de regras de pós-etiquetagem. In: *Anais do X Encontro de Linguística de Corpus – Aspectos metodológicos dos estudos de corpora***.** Belo Horizonte: UFMG, 2012.

ALENCAR, Leonel Figueiredo de. Novos recursos do Aelius para o processamento computacional raso do português. In: LAPORTE, E.; SMARSARO, A.; VALE, O. (Orgs.). *Dialogar é preciso: linguística para o processamento de línguas***.** Vitória: PPGEL/UFES, 2013.

ALENCAR, Leonel Figueiredo de; OTHERO, Gabriel de Ávila (Orgs.). *Abordagens computacionais da teoria da gramática*. Campinas: Mercado de Letras, 2011.

BERBER SARDINHA, Tony. O que é um corpus representativo? *DIRECT Papers 44*. São

Paulo / Liverpool: LAEL & AELSU, 2000a.

BERBER SARDINHA, Tony. Linguística de corpus: histórico e problemática. *D.E.L.T.A.*, v. 16, n. 2, 2000b.

BERBER SARDINHA, Tony. *Linguística de Corpus***.** Barueri, SP: Manole, 2004a.

BERBER SARDINHA, Tony. Linguística de Corpus: Uma entrevista com Tony Berber. *ReVEL*, v. 2, n. 3, 2004b.

DEZOTTI, Lucas Consolin. As “partes da oração” de Donato aos modistas. *ReVEL*, v. 8, n. 14, 2010.

GARSIDE, Roger; LEECH, Geoffrey; McENERY, Anthony. *Corpus annotation: linguistic information from computer text corpora*. London / New York: Longman, 1997.

HOLTZ, Louis. *Donat et la tradition de l’enseignement grammatical.* Paris: Centre National de la Recherche Scientifique, 1981.

RASO, Tommaso; MELLO, Heliana (Orgs.). *C-ORAL BRASIL I Corpus de referência de português brasileiro falado informal.* Belo Horizonte: Editora UFMG, 2012.

URBANO, Hudinilson. *Marcadores Conversacionais.* In: Análise de textos orais. São Paulo: Humanitás, 2010.