Avaliação da anotação semântica do PALAVRAS e sua pós-edição manual para o Corpus Summ-it

Élen Cátia Tomazela
etomazela@yahoo.com.br

Cláudia Dias de Barros
claudias84@gmail.com

Lucia Helena Machado Rino
lucia@dc.ufscar.br

Núcleo Interinstitucional de Linguística Computacional
Universidade Federal de São Carlos
São Carlos – SP, Brasil

Resumo

Este artigo apresenta uma avaliação da anotação semântica automática do parser PALAVRAS e sua pós-edição manual para um corpus de textos em português – o Corpus Summ-it. Essa pós-edição visou ao aprimoramento de um modelo linguístico para a sumarização automática de textos e buscou atribuir etiquetas semânticas mais adequadas aos itens lexicais, comparadas às empregadas pelo parser. Essa tarefa foi realizada por linguistas e os casos problemáticos são apresentados neste artigo, os quais levam a considerações sobre o próprio modelo de etiquetagem do PALAVRAS. O corpus revisado estará disponível para a comunidade e poderá ser útil para várias aplicações de Processamento de Línguas Naturais.

1 Introdução

Este artigo tem como finalidade explicitar a avaliação da anotação semântica provida pelo *parser* PALAVRAS (Bick 2000) para os textos que compõem o Corpus Summ-it (Collovini, Carbonel et al. 2007)[1] e o processo de pós-edição manual dessa etiquetagem. Esse corpus foi construído visando à sumarização automática de textos e é utilizado, particularmente, para a modelagem de critérios de decisão do sumarizador automático VeinSum (Carbonel 2007), cujo refinamento foi proposto em (Tomazela 2010).

O corpus possui vários tipos de anotação dos textos: as anotações morfossintática e semântica produzidas pelo *parser,* a anotação de cadeias de correferência (doravante CCRs) e a anotação retórica, esta na forma de estruturas RST (Mann & Thompson 1988). Somente as providas pelo PALAVRAS foram realizadas automaticamente; as demais são resultado de trabalho manual executado por especialistas nas devidas competências.

A anotação semântica é muito relevante para a melhoria do modelo do VeinSum porque é usada para especificar heurísticas de decisão para a escolha de segmentos textuais relevantes aos sumários, os quais são produzidos com foco no fenômeno do encadeamento referencial. Por isso, a anotação de CCRs contemplou os sintagmas nominais (aqui referidos por SNs) e, ainda, ao menos um que fosse expresso por uma descrição definida, por serem estas as construções de interesse para a correferenciação: para a sumarização baseada nas anotações semânticas, outras realizações linguísticas (p.ex., as pronominais) não seriam etiquetadas, a menos que a resolução anafórica fosse realizada automática e previamente, o que não ocorre no PALAVRAS.

O contexto que motiva o uso das anotações semânticas é descrito na Seção 2, o qual motivou a avaliação de desempenho do *parser* e a pós-edição de sua anotação. A descrição do modelo de anotação semântica automática se encontra na Seção 3, seguindo-se o relato das principais características do corpus e da metodologia empregada para a correção de suas etiquetas (Seção 4). Os principais problemas de etiquetagem semântica são descritos na Seção 5, seguindo-se a avaliação do desempenho do *parser* (Seção 6).

Neste artigo, anotação, etiquetagem e etiquetas semânticas são termos adotados para referência ao processo de marcação automática de itens lexicais com suas categorias semânticas, segundo o elenco de etiquetas fornecido pelo PALAVRAS, o que faz dele, além de um gerador de estruturas sintáticas, um *parser* ou etiquetador semântico. A pós-edição refere-se unicamente à revisão e correção das etiquetas semânticas atribuídas pelo sistema a todos os SNs correferentes pertencentes ao corpus.

[1] Disponível no Portal de Corpus do NILC, http://www.nilc.icmc.usp.br:8180/portal/.

2 *O modelo de sumarização automática do VeinSum*

O VeinSum é um sumarizador automático que segue a abordagem profunda, isto é, ele é baseado em processamento de conhecimento linguístico (Sparck-Jones 1999) para produzir sumários de textos-fonte[2]. Segundo essa abordagem, o sistema recorre a estruturas linguísticas cujos pressupostos teóricos servem para indicar a informação relevante para um sumário e sua organização textual. Essencialmente, essa organização refere-se à preservação da ordem original da informação e não há qualquer reescrita das unidades mínimas de significado tidas como relevantes para compô-lo. Ou seja, essas unidades são meramente *copiadas-e-coladas* do texto-fonte para o sumário. Está nessa etapa de reconhecimento de unidades relevantes para incluir em um sumário, portanto, o maior esforço do sistema para obter resultados satisfatórios. Como a unidade textual mínima é a sentencial, sentenças completas são copiadas nos sumários, resultando nos principais problemas de textualidade já descritos na literatura (p.ex., (Mani 2001)).

Propôs-se resolver no VeinSum um problema particular de textualidade: o de *clareza referencial*. Diz-se que um sumário apresenta clareza referencial quando não há *quebras* de CCRs. Uma quebra de CCR, por sua vez, ocorre quando não é possível, ao leitor, identificar a quem ou a que um determinado pronome ou SN está se referindo (definição da DUC2005 – *Document Understanding Conference*)[3]. Assim, a meta principal do sistema é produzir sumários automáticos que sejam claros referencialmente.

Embora a garantia de textualidade envolva critérios intra e extralinguísticos (Beaugrande & Dressler 1981) e a própria definição de clareza referencial explicite a intervenção do leitor, a modelagem do sumarizador automático contempla somente o nível intratextual, para evitar os demais problemas da referenciação, os quais, até o momento, são intratáveis computacionalmente. Como resultado, somente o aspecto coesivo é considerado, fugindo do escopo da abordagem quaisquer outras considerações relativas à coerência textual, como as apontadas em (Halliday & Hasan 1976), (Marcuschi 1983) ou (Koch & Travaglia 2004). Logo, tratar da clareza referencial para gerar um sumário consiste em determinar automaticamente qual é a sentença com maior probabilidade de conter o antecedente mais completo de um componente anafórico já incluído no sumário.

Uma quebra de clareza referencial é evidenciada no sumário a seguir, gerado automaticamente para o texto-fonte CIENCIA_2001_6410, dado como entrada ao sistema[4]. Esse sumário contém a expressão anafórica '**o pesquisador**' sem que seu antecedente esteja explícito. O excerto do texto-fonte em que essa anáfora se insere segue o sumário. Nota-se que a menção ao antecedente se encontra na sentença imediatamente anterior, a qual foi desconsiderada pelo sumarizador automático, em sua decisão do que incluir no sumário.

> Ao contrário do que muita gente pensa, a internet não está reduzindo os contatos entre as pessoas nem substituindo-os por relações impessoais conduzidas por computador. Segundo **o pesquisador**, os contatos via redes de computadores estão na verdade ampliando a socialização das pessoas.

Sumário do texto CIENCIA_2001_6410

> Ao contrário do que muita gente pensa, a internet não está reduzindo os contatos entre as pessoas nem substituindo-os por relações impessoais conduzidas por computador. A conclusão é de **Barry Ellman**, do Centro para Estudos Urbanos e Comunitários da Universidade de Toronto, Canadá. Segundo **o pesquisador**, os contatos via redes de computadores estão na verdade ampliando a socialização das pessoas.

Excerto do texto-fonte CIENCIA_2001_6410

Para tratar a clareza referencial de fato seria necessário que o sistema computacional identificasse as CCRs e apontasse seus componentes que resolvem as referências, quer elas sejam anafóricas, quer sejam de qualquer outro tipo descrito na literatura (Coelho, Muller et al. 2006). Esse é o problema que as iniciativas de

[2] A única acepção adotada aqui, para o termo *sumário*, é a de *resumo da fonte de informação.*

[3] http://duc.nist.gov/duc2005/.

[4] Todos os textos ilustrados neste artigo foram extraídos do Corpus Summ-it e seus sumários automáticos, gerados pelo VeinSum.

resolução anafórica automática pretendem resolver. No entanto, as soluções computacionais são aproximações que frequentemente carecem de qualidade, em geral porque os resolvedores anafóricos não conseguem tratar adequadamente esse fenômeno linguístico, já de natureza complexa, que demanda modelos de resolução automática incompletos ou inexatos.

No projeto do VeinSum, optou-se por manter o foco somente na questão de sumarização, evitando aumentar sua complexidade com a agregação de um módulo de resolução anafórica, muito embora a ausência desse processo seja, reconhecidamente, um dos maiores entraves para a Sumarização Automática (Mitkov 1998; Cristea, Postolache et al. 2003) e, em geral, para os sistemas de PLN[5] (Mitkov 2002; Chaves 2007).

A proposta alternativa para buscar a clareza referencial foi a de fazer o sistema manipular as estruturas RST dos textos-fonte. Assim, qualquer texto a sumarizar é, primeiramente, estruturado retoricamente e é a partir de sua estrutura RST que se busca determinar quais as unidades textuais a incorporar aos sumários.

Além de não resolver anáforas explicitamente, o VeinSum sequer é capaz de detectar os termos anafóricos. Na verdade, ele procura delimitar os contextos de possíveis unidades correferentes (os quais incluem as possíveis anáforas e seus antecedentes) somente com base nas estruturas RST, ou seja, na sua posição nas árvores dos textos-fonte. Essa delimitação dos contextos correferenciais fica a cargo da Teoria das Veias, ou VT (Cristea, Ide et al. 1998). Associada à RST, ela o faz com base no *domínio de acessibilidade referencial* (doravante, *acc*) de cada unidade textual da árvore.

O *acc* é, assim, o conjunto de todas as unidades que possam fazer parte da CCR de uma unidade anafórica, a qual também é incluída nesse conjunto. Na ausência da resolução anafórica como tal, o *acc* se constitui, portanto, das sentenças do texto-fonte que, hipoteticamente, são correferentes. Esse é o ponto de partida do VeinSum para buscar manter a clareza referencial dos sumários.

O problema do sistema pode ser descrito, portanto, como o problema de se reconhecer, dentre as N unidades textuais que compõem um texto-fonte e que se encontram relacionadas em sua estrutura RST, quais são as M unidades (M menor que N) que comporão o sumário correspondente, sem que haja quebra da clareza referencial. A VT, juntamente com a RST, fornece todos os *acc*s das sentenças do texto-fonte.

Para indicar quais as M sentenças que serão escolhidas, agrega-se aos dois modelos anteriores o Modelo de Saliência (Marcu 2000), que indica a classificação de saliência das N unidades a partir da qual as M unidades são escolhidas. Ante as restrições de saliência, clareza referencial e taxa de compressão (restrição fundamental da sumarização automática), que são consideradas em conjunto, o sistema finalmente produz o sumário integral.

Um dos motivos da fragilidade dos resultados do VeinSum, como o ilustrado pelo sumário do texto CIENCIA_2001_6410, é que, ao ter que obedecer à taxa de compressão, se necessário o sistema relaxa a restrição de saliência, desprezando sentenças mais salientes para manter integralmente os *acc*s de unidades já escolhidas para compor o sumário. Com isso, informações mais importantes do texto podem ser desprezadas, prejudicando a qualidade do sumário, quando comparado ao seu texto-fonte.

O que gerou a proposta de refinamento de Tomazela (2010) foi a observação de que os *acc*s também poderiam ser reduzidos, pois os contextos de prováveis unidades correferentes apontados pela VT não asseguram, de fato, quais delas são essenciais para a clareza referencial. No melhor caso, bastaria manter, do *acc*, as sentenças que contêm a anáfora e a que contém seu antecedente mais completo.

Assim, na tentativa de reduzir os *acc*s, propôs-se o uso de informações semânticas provindas da anotação do PALAVRAS como coadjuvante dos modelos descritos. Supôs-se, nesse caso, que o problema não estaria na estruturação RST, nem na determinação dos *acc*s de cada componente textual, muito embora tanto a RST quanto a VT tragam reconhecidos problemas para a manipulação de segmentos textuais (Cristea, Postolache et al. 2005; Carbonel 2007; Tomazela & Rino 2009).

Em linhas gerais, buscando selecionar menos sentenças de cada *acc*, o novo sumarizador procede da seguinte forma: uma vez escolhida uma sentença para compor o sumário, as etiquetas semânticas dos núcleos de cada um de seus SNs são usadas para buscar o provável antecedente de uma anáfora hipotética dessa sentença. Esse é apontado como a unidade do *acc* que contenha um ou mais SNs com maior similaridade semântica com os núcleos dos SNs da unidade já escolhida.

É, portanto, a similaridade semântica entre componentes de várias sentenças apontadas no

[5] **P**rocessamento automático de **L**ínguas **N**aturais.

acc que irá indicar a possibilidade de manter a clareza referencial no sumário e, ao mesmo tempo, permitir que a classificação das unidades salientes seja respeitada, para melhor aproximação com a preservação das informações mais relevantes do texto-fonte.

O problema recai, portanto, em como distinguir componentes mais similares – aqueles que possam indicar uma ligação forte de correferência. Isso é feito traçando-se a relação entre as etiquetas semânticas fornecidas pelo PALAVRAS, para os SNs em foco, isto é, os SNs que possivelmente sejam correferentes,. Com base nessa ideia, é que se buscou definir heurísticas para a escolha das unidades relevantes que atendessem aos critérios de similaridade semântica, ditados por um modelo de similaridade baseado na distribuição das etiquetas num corpus (Tomazela 2010). Esse modelo é descrito na próxima seção.

3 O modelo de anotação semântica do PALAVRAS

O processamento semântico do PALAVRAS visa à atribuição de uma etiqueta semântica que indique, *aproximadamente,* o significado de cada item lexical de um texto. Para isso, não se consideram modelos clássicos de semântica lexical, nos quais se buscam significados através de definições dicionarizadas ou por uma classificação ontológica, mas sim, combinações de traços semânticos, os quais fornecem uma identidade ao item lexical. Essa anotação conta com 215 etiquetas semânticas e se baseia em 16 traços, os quais supostamente representam o contexto semântico de quaisquer conceitos usados na produção de uma mensagem (Bick 2000). Note-se que essa concepção implica considerar o modelo semântico independente de língua natural.

Nesse modelo de classificação, são considerados somente os substantivos, entidades nomeadas e alguns adjetivos, para os quais é possível atribuir um valor semântico. As entidades nomeadas, neste trabalho, são o mesmo que entidades mencionadas (Santos 2007), denotadas por nomes próprios que podem indicar nomes de pessoas, organizações, acontecimentos, locais, coisas, obras e conceitos abstratos.

A identificação de itens lexicais similares é atribuída à chamada similaridade prototípica, a qual permite colocar em contexto de uso a configuração semântica, sem que se necessite de coincidências absolutas de significado. Essa medida de similaridade de cada item lexical é proporcional ao número de traços semânticos que compartilham: Bick supõe que, quanto maior esse número, mais similares são os itens lexicais. Daí a possibilidade de agregar, em um único conjunto ou, no caso de interesse para o VeinSum, em uma única heurística, etiquetas semânticas que indiquem itens lexicais possivelmente correferentes.

Foi essa ideia de similaridade semântica baseada nas etiquetas do PALAVRAS que motivou a proposta de se definirem heurísticas para a sumarização automática de textos em português. Porém, ao se analisar a anotação semântica do Corpus Summ-it, descobriram-se vários casos de inadequação da etiquetagem automática, residindo aí a motivação para a sua pós-edição manual e consequente avaliação do *parser* apresentadas neste artigo. O Corpus Summ-it foi o instrumento central para a engenharia do conhecimento visando à formalização de todo o processo.

4 O Corpus Summ-it

O Corpus Summ-it configura-se como o primeiro corpus anotado manualmente com CCRs para textos jornalísticos em português. Foi construído para atender a diversos interesses, dentre os quais os de pesquisa e desenvolvimento de sistemas de sumarização automática de textos, uma das principais áreas de pesquisa do NILC[6]. É composto de 50 textos do caderno de Ciências da Folha de São Paulo, cada um deles de tamanho que varia de 27 a 654 palavras(1/2 a 1 ½ página em formato A4). Os textos contêm de 3 a 24 CCRs, totalizando 589 CCRs no corpus todo. A CCR mais longa contém 16 SNs e a mais curta, apenas 2.

A importância da anotação semântica para o VeinSum se deve ao fato de ele usar os *accs* como conjuntos indicativos do contexto de ocorrência de CCRs e estas terem seus SNs já anotados semanticamente. Isso permite elaborar o processo de determinação dos segmentos a compor um sumário proposto como melhoria do sistema. O fato de os *accs* serem derivados das estruturas RST dos textos-fonte justifica a existência da anotação RST do corpus todo. Entretanto, para o novo processo de minimização dos *accs* ocorrem dois entraves: não se sabe qual o SN anafórico, tampouco qual o SN que poderia ser seu antecedente, daí a busca de heurísticas que possam indicar possíveis contextos correferentes pelas etiquetas semânticas dos componentes dos

[6] Núcleo Interinstitucional de Linguística Computacional, sediado em São Carlos, SP - http://www.nilc.icmc.usp.br/nilc/index.html.

accs. Também por essa razão e pela verificação de problemas na etiquetagem automática, originou-se a necessidade de se revisar os resultados do PALAVRAS. Dessa forma, tentou-se garantir a confiabilidade das heurísticas a incorporar ao VeinSum.

4.1 A necessidade de revisão do corpus

Como o foco é simplesmente a minimização dos *accs,* a pós-edição do corpus Summ-it se restringiu aos SNs que aparecem nas CCRs. Particularmente, foram analisados os substantivos desses SNs, já que eles são os únicos que contêm etiquetas semânticas expressivas, como já mencionado.

Nesta seção relatam-se os principais desvios de anotação semântica do PALAVRAS para os itens lexicais em questão. Foram identificados três problemas significativos: o de segmentação, o de etiquetagem, e o de desambiguação das etiquetas semânticas. Certamente esses problemas são interdependentes: a má segmentação textual interfere nos demais. A desambiguação de sentido é, na verdade, um problema da própria etiquetagem: etiquetas equivocadas podem ser atribuídas por não haver uma determinação clara (ou menos problemática) do significado de algum componente textual. Mesmo a anotação morfossintática (em inglês, *POS tagging*) depende da segmentação, por um lado, e interfere no desempenho semântico, por outro. Ou seja, uma má segmentação textual constitui o primeiro entrave para os demais processos, fato amplamente reconhecido na área de PLN (Pardo & Nunes 2002). Sobretudo no caso de CCRs, julgou-se que a segmentação inadequada pode corromper o encadeamento referencial e levar a problemas sérios para que o modelo de decisão do VeinSum assegure a clareza referencial.

4.2 Os pressupostos para a revisão do corpus

Primeiramente, para manter a tarefa de pós-edição consistente, determinou-se que ela seria feita por duas linguistas especialistas no elenco de etiquetas semânticas do PALAVRAS[7] e que a concordância em suas decisões seria assegurada (Pressuposto 1). Desse modo, o corpus Summ-it revisado pode servir à engenharia de conhecimento na Sumarização Automática (para avaliação ou validação de sistemas), mas também a tarefas que não as de PLN, como as de Linguística de Corpus.

Adotou-se por base as definições originais das etiquetas, o que levou à limitação das possíveis acepções a suas denotações fixas (Pressuposto 2). Esse método de análise semântica está em consonância com as instruções usuais da MUC (*Message Understanding Conference*), ao contrário do que é sugerido em (Santos & Cardoso 2007).

Um exemplo claro da aplicação dessa estratégia no processo de revisão refere-se ao uso metonímico de termos que indicam localizações[8]: no Summ-it, há ocorrências de 'Brasil', cuja etiqueta semântica categorial (ou denotacional) é **<Lciv>** (*Civitas, town, country, county, cidade, país*). Entretanto, o PALAVRAS atribui a esses termos, muitas vezes, a etiqueta **<inst>** *(institution*), claramente metonímica. Todos esses casos foram alterados para <Lciv>.

O terceiro pressuposto foi o de que a revisão em curso não infringiria os próprios pressupostos do PALAVRAS, de que protótipos semânticos podem ser usados para indicar a similaridade (ou dissimilaridade) entre vários itens lexicais. Ao contrário, seria possível usar a prototipagem para fundamentar a revisão – e, certamente, para traçar mecanismos de sumarização automática (Pressuposto 3).

Finalmente, perseguiu-se a perspectiva de que o PALAVRAS se destina ao processamento do português e, assim, tem seu elenco de etiquetas igualmente aplicável e reusável para o processamento dos textos nessa língua, os quais são os objetos de interesse para a sumarização automática em curso.

Vale notar que, exceto pelo uso do elenco de etiquetas semânticas do PALAVRAS, cuja dependência de qualquer língua-fonte pode ser questionada, as demais linguagens de representação adotadas nas anotações do corpus (estruturas RST e estruturas de veias, no caso) são independentes de língua natural.

4.3 A preparação do corpus e a metodologia de revisão

Para evitar que constantes atualizações do PALAVRAS prejudicassem a consistência da tarefa de revisão manual das etiquetas semânticas,

[7] Acessível pela Internet: (http://beta.visl.sdu.dk/visl/pt/info/portsymbol.html#semtags).

[8] Conforme discussão de Santos (2007).

adotou-se sua versão de fevereiro de 2007 e, assim, o elenco das 215 etiquetas semânticas, juntamente com suas definições, foi mantido constante[9].

Foram utilizadas diretamente as saídas do sistema para cada um dos 50 textos do corpus: arquivos XML. Esses dados de cada texto foram agrupados em uma única planilha Excel, a qual consistiu o material de trabalho das especialistas linguistas. Como já mencionado, restritos os SNs aos componentes de CCRs, para cada texto somente as anotações semânticas desses dados constam da planilha. As correções das etiquetas foram inseridas também nesse arquivo, de forma que toda síntese numérica (totalizações de casos, estatísticas de ocorrência, etc.) necessária para a análise foi produzida automaticamente, via programação no próprio ambiente da planilha Excel. A partir desse processamento foi possível avaliar o desempenho do PALAVRAS para textos isolados ou em conjunto, resultando na síntese apresentada na Seção 5.

5 A pós-edição do Corpus Summ-it

Considerando a interdependência entre a segmentação e os demais processos do PALAVRAS, relatam-se aqui primeiramente os casos problemáticos de segmentação, para depois apresentar-se os problemas de etiquetagem semântica, propriamente ditos. Maiores detalhes dessa tarefa podem ser encontrados em (Tomazela & Rino 2010).

5.1 Problemas de segmentação

Os casos mais problemáticos de segmentação textual do PALAVRAS residiram na confusa identificação de lexias complexas e de entidades nomeadas. Várias lexias complexas foram consideradas lexias simples, ou seja, foram processadas em componentes separados, com o desmembramento de uma única entidade em vários SNs. Esse padrão foi identificado para as lexias compostas de 'substantivo + adjetivo', como nos seguintes exemplos do corpus:

- 'vaso sanguíneo', sendo 'vaso' etiquetado com **<con>** (*container*) e 'sanguíneo' ignorado. A lexia deveria ser etiquetada com **<an>** (*anatomical noun, umbrela tag - carótida, dorso).*
- 'cadeia evolutiva', sendo 'cadeia' etiquetada com **<inst>** (*institution),* em vez de **<ax>** (Abstract/concept, neither countable nor mass – *endogamia)* e 'evolutiva' ignorada;
- 'batimento cardíaco', sendo 'batimento' etiquetado com **<act>** (*Action, umbrella tag - +CONTROL, PERFECTIVE)*, em vez de **<process>** (process -CONTROL, -PERFECTIVE, cp. <event>, *balcanização, convecção, estagnação*) e 'cardíaco' ignorado.

Esses exemplos evidenciam que haverá prejuízo para a identificação de termos correferentes com a etiquetagem independente: os adjetivos ignorados é que realmente determinam o significado das lexias complexas.

Opostamente a esses casos, a ferramenta aglutinou vários SNs em uma única entidade nomeada, conforme os seguintes exemplos:

- Pesquisadores do Museu Nacional do Rio de Janeiro

  Aqui tem-se o SN 'Pesquisadores' e as entidades nomeadas "Museu Nacional" e "Rio de Janeiro'. O *parser* atribui a etiqueta **<hum>** (*person name)* para todo esse trecho, pois o considera uma única entidade. No entanto, três etiquetas distintas deveriam ter sido atribuídas: 'Pesquisadores', com etiqueta **<Hprof>** (Professional human – *marinheiro*), 'Museu Nacional", com **<org>** (*commercial or non-commercial, non-administrative, non-party organisations*) e 'Rio de Janeiro', com **<civ>** (*civitas - country, town, state, cp. <Lciv>*).

- Organização das Nações Unidas

  O *parser* etiquetou separadamente os seguintes itens lexicais: 'Organização', com **<np-close>**, cuja definição não é encontrada no elenco de etiquetas; 'Nações', com **<HH>** (*Group of humans - organisations, teams, companies, e.g. editora)* e 'Unidas' não recebeu etiqueta semântica alguma. Caso essa entidade nomeada não fosse desmembrada, sua etiqueta deveria ser **<org>.**

Considerou-se que, ao não se reconhecer entidades nomeadas de um texto, a proposta de identificação de elementos correferentes por suas etiquetas semânticas se tornaria mais difícil.

[9] Embora essa preocupação seja procedente, o elenco permanece o mesmo até a presente data.

Entretanto, esta suposição merece uma investigação mais profunda no futuro.

No total, foram identificados 104 casos problemáticos de segmentação (vide seção 5.3), os quais incluem a identificação de lexias complexas e entidades nomeadas.

5.2 Problemas de etiquetação

Primeiramente apresentam-se alguns detalhes sobre o procedimento de verificação da etiquetagem, para depois relatarem-se alguns casos pitorescos do corpus.

5.2.1 Especificidades da revisão

De forma geral, qualquer correção de etiquetas atribuídas pelo sistema aos itens lexicais somente se deu quando a etiqueta apresentou desvios semânticos consideráveis. Nesse caso, optou-se por utilizar etiquetas mais específicas sempre que possível. Entretanto, as etiquetas genéricas produzidas pelo sistema foram mantidas sempre que julgadas apropriadas, buscando não penalizar excessivamente a avaliação de desempenho pretendida. Ou seja, somente foram alterados os casos em que ou a etiqueta era claramente indevida, por ser conflitante com os traços semânticos do componente lexical, ou a etiqueta era tão específica que não correspondia ao seu significado adequado. Nesse caso, adotou-se uma etiqueta referente a um conceito mais geral. Exemplos disso ocorrem para os itens lexicais 'bicho' e 'animal', ambos etiquetados com **<Azo>** (*land animal).* Considerou-se essa etiqueta restritiva porque as acepções desses itens lexicais no corpus em foco abrangem também animais aquáticos. A etiqueta mais genérica atribuída foi, portanto, **<A>** (*Animal, umbrella tag - clone, fêmea, fóssil, parasito, predador)*, a qual, na existência de uma ontologia apropriada, seria considerada um hiperônimo da etiqueta **<Azo>**.

Considerou-se o contexto de ocorrência dos itens lexicais e, assim, recorreu-se aos textos-fonte correspondentes, sobretudo quando se necessitou interpretar itens lexicais anafóricos cujos referentes não estavam acessíveis na planilha Excel.

Também foi necessário verificar os casos de delimitação das entidades nomeadas, para atribuir-lhes uma única etiqueta a partir de sua análise como um todo (afinal, a semântica de um componente desse tipo não é a soma da semântica de suas partes).

Quando não se conseguiu definir a melhor etiqueta para corrigir a automática, recorreu-se ao tópico (ou assunto) do texto, para traçar seu interrelacionamento. Por exemplo, a menção anafórica 'os pesquisados' em uma certa CCR pode se referir a pessoas, animais, medicamentos ou produtos. A partir do tópico principal do texto em que está inserida (CIENCIA_2000_17101), expresso pelo segmento "a alteração da Declaração de Helsinque, na qual os cientistas não se obrigariam a fornecer aos doentes o melhor tratamento conhecido para uma doença", é possível determinar que esse SN se refere a 'os doentes' e, portanto, deve ser etiquetado com **<H>** (*human, umbrella tag).*

Mediante esses casos, vale lembrar que o *parser* não propõe fazer resolução anafórica e, por isso, não tem obrigação de reconhecer esses antecedentes. Porém, ao não fazê-lo, produz etiquetas que podem trazer problemas à clareza referencial dos sumários.

Analisou-se ainda o aspecto dos itens lexicais, particularmente quando indicavam eventos, ações, atividades ou processos. Nesses casos há etiquetas específicas que distinguem a valência (+/-) do traço semântico PERFECTIVE: +PERFECTIVE indica conceito pontual; –PERFECTIVE, conceito progressivo. Distinguiram-se, também, as valências do traço semântico CONTROL, isto é, se os conceitos apresentados eram passíveis ou não de serem controlados. As etiquetas que tratam desses casos são indicadas abaixo:

- **<activity>** (*Activity, umbrella tag - +CONTROL, IMPERFECTIVE, correria, manejo);*
- **<act>** (*Action, umbrella tag - +CONTROL, PERFECTIVE);*
- **<event>** (*event, -CONTROL, PERFECTIVE, milagre, morte)*;
- **<process>** (*process, -CONTROL, –PERFECTIVE, cp. <event>, balcanização, convecção, estagnação)*[10].

Caso as estratégias relativas ao contexto de ocorrência e às definições das etiquetas ainda não fossem suficientes para determinar etiquetas apropriadas, recorreu-se à WordNet (Fellbaum 1998), para buscar seus traços semânticos.

5.2.2 Ocorrências problemáticas no Corpus Summ-it

Destacam-se, aqui, alguns dos problemas de etiquetação mais significativos no corpus:

[10] Entende-se' –PERFECTIVE' como 'IMPERFECTIVE', neste caso.

- Nomes científicos, muito presentes nos textos do corpus em uso, quase sempre são etiquetados ou segmentados erroneamente. '*Tyrannossaurus rex*', p.ex., é etiquetado com **<inst>**, quando deveria receber a etiqueta **<meta>** (meta noun - *tipo*, *espécie*)**.**

- 'células-tronco', quando inicia a oração, recebe etiqueta **<Acell>** (*Cell-animal - bacteria, blood cells: linfócito)*; quando ocorre intraoracionalmente, recebe etiqueta **<HH>** (*Group of humans - organisations, teams, companies, e.g. editora)*, o que contradiz o fato de serem correferentes, já que '*animal celular*' não pode ser correferente a um '*grupo de humanos*'.

- Apesar de sinônimos, alguns itens lexicais correferentes apresentam etiquetas diferentes, como: 'cachorro' **-** etiquetado com **<Azo>** (*Land-animal - raposa)* e 'cão' etiquetado com **<Adom>** (*Domestic animal or big mammal - terneiro, leão/leoa, cachorro)*. O que justifica o fato de 'cachorro' ser animal terrestre e 'cão' ser animal doméstico não é claro.
- Caso análogo ocorre com 'CO2', que recebe etiqueta **<cm-chem>** (*chemical substance, also biological - acetileno, amônio, anilina, bilirrubina)* e 'gás carbônico', etiquetado com **<mat>** (material - *argila, bronze, granito*, cf. <cm>).
- O item lexical 'atmosfera' recebe etiquetas diferentes dependendo da palavra que o segue, como em: 'atmosfera da Terra' e 'atmosfera terrestre', com etiquetas **<Ltop>** (*Geographical, natural place - promontório, pântano)* e **<sit>** (*psychological situation or physical state of affairs - reclusão, arruaça, ilegalidade, more complex & more "locative" than <state> & <state-h>)* respectivamente.

Esses exemplos sugerem que não há tratamento de sinonímia no PALAVRAS, o que também compromete o modelo de busca de itens correferentes. Eles constituem alguns dos exemplos mais problemáticos observados na revisão. A desambiguação de itens lexicais também se mostrou frágil.

5.2.3 A etiquetagem de itens ambíguos

Para determinar o significado adequado, vários fatores entram em perspectiva, sendo dos mais significativos o contexto de ocorrência do item lexical. Se o modelo semântico do *parser* pretende apontar as etiquetas semânticas aproximadas para componentes textuais, ele deveria prover mecanismos para tratar esses fenômenos. Um exemplo dessa deficiência ocorre com o item lexical 'clone', com etiqueta **<H>,** a qual somente se refere a clones humanos. No entanto, os contextos de ocorrência desse item no corpus mostram que esse termo se aplica a clones de animais e, assim, a etiqueta utilizada deveria ser a mais genérica **<A>**.

Já a desambiguação de vários itens lexicais em SNs compostos seria beneficiada se seu interrelacionamento fosse considerado na determinação do significado. O PALAVRAS não parece considerar esse contexto de ocorrência, como ilustram os exemplos a seguir:

- 'as patas e bacia do animal', em que 'bacia' recebe etiqueta **<con>** (*container),* quando deveria receber **<anmov>** (*Movable anatomy - arm, leg, braço, bíceps, cotovelo);*
- 'a física nuclear Eva Maria', em que 'física' recebe etiqueta **<domain>** (*subject matter, profession, cf. <genre>, anatomia, citricultura, datilografia)*, quando deveria ser **<Hprof>** (Professional human - *marinheiro*, also sport, hobby - *alpinista*)*;*
- 'populações de pinguins', em que 'populações' recebe etiqueta **<HH>** (*Group of humans - organisations, teams, companies, e.g. editora),* em vez de **<AA>** (*Group of animals - cardume, enxame, passarada, ninhada*)*;*
- 'esqueleto do navio', em que 'esqueleto' recebe etiqueta **<Hmyth>** (*Humanoid mythical - gods, fairy tale humanoids, curupira, duende),* em vez de **<part-build>** (structural part of building or vehicle - *balustrada, porta, estai*).
- 'filhote' é etiquetado com **<H>** (*Human, umbrella tag)* quando o sentido de animal – etiqueta **<A>** – indicado pelo contexto é ignorado.

Esses exemplos evidenciam a necessidade de um tratamento automático mais elaborado para os casos que envolvem aspectos contextuais.

5.3 Síntese da pós-edição manual do corpus

A Tabela 1 mostra os dados gerais de correção do corpus ('SUBSTs', aqui, é limitado aos substantivos de SNs presentes nas CCRs do corpus). A média de correções de etiquetas semânticas no corpus foi de 41%. A porcentagem de erros de segmentação foi de 4%. Essa baixa porcentagem demonstra que eventuais problemas de etiquetação morfossintática ou semântica não foram causados significativamente pela segmentação automática do PALAVRAS, no corpus Summ-it. Não foi analisada isoladamente a influência da etiquetagem morfossintática na etiquetagem semântica.

Do elenco total de etiquetas (215), somente 115 ocorreram no Corpus Summ-it, segundo a revisão manual aqui relatada. Elas são reproduzidas na Tabela 2.

No tocante aos pressupostos desse trabalho, essa revisão constitui somente o passo inicial para se verificar a adequação da estratégia a outros corpora e, assim, a consistência da revisão aqui apresentada, reafirmando o Pressuposto 1. A limitação da revisão a denotações fixas (Pressuposto 2) certamente é um fator limitante. Porém, considerando-se a perspectiva de se ter um modelo automático, ela representa uma decisão razoável a se adotar, corroborada, inclusive, pelas diretrizes da MUC.

Entretanto, a questão mais polêmica sugerida pela análise aqui descrita diz respeito ao Pressuposto 3, isto é, à incorporação da ideia de protótipos semânticos que propiciem o reconhecimento de entidades similares ou dissimilares. Garantir isso pareceu impossível, dada a especificidade da classificação proposta por Bick (inclusive o fato de ela se basear em corpora de textos), ao fato de ela se inserir no contexto de tradução automática e, até, à necessidade, em alguns casos, de se buscar os vínculos em contexto para se determinar as etiquetas mais adequadas.

Particularmente, buscar a base teórica para a definição dos protótipos semânticos do *parser* foi uma tarefa difícil. Mesmo a forma como Bick propõe obter as categorias baseadas nos protótipos não está clara: foi feita com base em corpora, visando especialmente a tradução automática, com o norueguês e o inglês como línguas interagentes. Entretanto, o *parser* está disponível para anotação de textos em português.

Essas limitações pareceram bastante severas para o reuso das etiquetas e também para a interpretação de sua definição ante a tarefa de revisão. Vale ressaltar que a opção de se escolher sempre uma etiqueta mais genérica (opção plausível em diversas aplicações) não esteve em foco porque ela não permitira alcançar o objetivo de distinguir elementos correferentes. Além disso, o contexto de sumarização automática em foco pode introduzir variações do contexto original ou, até, ser inadequado para se buscar similaridades semânticas pela diferenciação de equivalentes de tradução (Santos 1990). Esta é uma questão ainda em aberto.

Por fim, a geração das heurísticas baseadas na idéia de proximidade semântica das etiquetas que pudessem indicar elementos correferentes foi dificultada porque não foi possível mediante o Pressuposto 3, traçar uma relação clara com a idéia de prototipagem semântica pelo reconhecimento de equivalentes. Esta questão não está em foco neste texto, mas é abordada em (Tomazela 2010).

Ressalta-se que as heurísticas foram geradas somente depois da pós-edição manual da etiquetagem semântica porque, se assim não fosse (isto é, se elas fossem geradas a partir do corpus diretamente anotado pelo PALAVRAS), elas seriam obviamente inválidas e não serviriam ao propósito deste trabalho, pois não assegurariam a indicação de possíveis itens lexicais correferentes.

6 Avaliação do desempenho do PALAVRAS

Dentre as principais dificuldades encontradas no processo de correção das etiquetas semânticas estão: i) a atribuição de etiquetas para itens lexicais de domínios específicos do conhecimento; ii) a inadequação das definições das etiquetas e de seus exemplos, presentes no PALAVRAS; iii) o reconhecimento de etiquetas muito genéricas, muito específicas ou ainda muito abstratas; iv) a dificuldade de adequação de um item lexical a uma única etiqueta, já que muitos deles podem ser etiquetados de várias formas.

O caso (i) foi particularmente complicado, pois, apesar de o corpus ser de domínio geral, há textos de assuntos muito particulares para algumas áreas da ciência. Para esses, o conhecimento especialista foi crucial e as linguistas precisaram recorrer a especialistas das áreas em foco, para determinar as etiquetas que melhor refletissem a natureza dos itens lexicais.

Texto-fonte	# SUBSTs	# SUBSTs corrigidos	# erros de segmentação	% Correção das etiquetas
CIENCIA_2005_6507	24	16	2	66.67%
CIENCIA_2003_6465	41	27	4	65.85%
CIENCIA_2003_24212	106	65	4	61.32%
CIENCIA_2001_19858	63	38	6	60.32%
CIENCIA_2005_28752	72	43	2	59.72%
CIENCIA_2001_6423	17	10	2	58.82%
CIENCIA_2001_6410	27	15	4	55.56%
CIENCIA_2000_17088	62	34	1	54.84%
CIENCIA_2002_22029	99	52	1	52.53%
CIENCIA_2002_6441	21	11	0	52.38%
CIENCIA_2000_6381	60	31	1	51.67%
CIENCIA_2000_17113	76	39	1	51.32%
CIENCIA_2005_28764	98	50	0	51.02%
CIENCIA_2000_17108	55	28	1	50.91%
CIENCIA_2000_6389	31	15	0	48.39%
CIENCIA_2004_26417	52	25	7	48.08%
CIENCIA_2005_28755	82	38	2	46.34%
CIENCIA_2000_17101	59	27	1	45.76%
CIENCIA_2002_22023	60	27	1	45.00%
CIENCIA_2005_28754	65	29	4	44.62%
CIENCIA_2002_22015	70	31	3	44.29%
CIENCIA_2004_6480	50	22	0	44.00%
CIENCIA_2003_24226	84	36	3	42.86%
CIENCIA_2005_28756	75	31	0	41.33%
CIENCIA_2001_6414	30	12	1	40.00%
CIENCIA_2004_26415	33	13	1	39.39%
CIENCIA_2005_28766	107	42	8	39.25%
CIENCIA_2002_22027	91	35	1	38.46%
CIENCIA_2000_17082	37	14	1	37.84%
CIENCIA_2000_17109	75	28	1	37.33%
CIENCIA_2004_6494	30	11	7	36.67%
CIENCIA_2005_6515	41	15	0	36.59%
CIENCIA_2003_6472	22	8	0	36.36%
CIENCIA_2005_28774	85	30	0	35.29%
CIENCIA_2000_17112	54	18	6	33.33%
CIENCIA_2001_6406	21	7	0	33.33%
CIENCIA_2005_6514	37	12	0	32.43%
CIENCIA_2004_26423	115	37	10	32.17%
CIENCIA_2005_6518	45	14	0	31.11%
CIENCIA_2004_6488	13	4	0	30.77%
CIENCIA_2001_6416	43	13	1	30.23%
CIENCIA_2000_6391	41	12	2	29.27%
CIENCIA_2000_6380	31	9	0	29.03%
CIENCIA_2005_28747	42	12	4	28.57%
CIENCIA_2004_26425	99	23	1	23.23%
CIENCIA_2003_24219	81	17	4	20.99%
CIENCIA_2002_22005	62	12	4	19.35%
CIENCIA_2002_22010	36	6	0	16.67%
CIENCIA_2003_6457	45	6	2	13.33%
CIENCIA_2005_28743	35	1	0	2.86%
TOTAIS	**2800**	**1151**	**104**	**2027.46%**

Tabela 1 – Quadro geral de correção da anotação semântica do Corpus Summ-it

A	Aorn	coll-cc	Hbio	mat	sick
AA	Azo	coll-sem	Hfam	meta	sick-c
absname	B	con	HH	mon	sit
ac	BB	conv	Hideo	month	site
ac-cat	build	cord	Hnat	object	suborg
Acell	Bveg	dir	Hprof	occ	temp
ac-sign	cc	disease	Hsick	org	therapy
act	cc-board	domain	hum	part	tool
act-d	cc-fire	drink	inst	part-build	tube
activity	cc-r	dur	L	party	unit
act-s	cc-rag	event	Labs	per	V
admin	cc-stone	f	Lciv	percep-w	Vair
Adom	civ	f-c	Lcover	pict	virtual
Aent	cm	f-h	Lh	piece	VV
am	cm-chem	food	ling	plan	Vwater
amount	cm-gas	food-h	Lopening	process	
an	cm-liq	f-q	Lstar	pub	
anbo	cm-rem	fruit	Lsurf	sem-c	
anmov	col	H	Ltop	sem-r	
anorg	coll	Hattr	Lwater	sem-s	

Tabela 2 – Etiquetas ocorrentes no corpus

O caso (ii) levou a uma grande dificuldade para a análise semântica, pois nem os exemplos fornecidos com o elenco de etiquetas foram suficientes para deixar claras muitas das definições. Etiquetas diferentes destinam-se a designar objetos semânticos diferentes, porém, quando se analisam os exemplos que acompanham suas definições, elas não parecem se diferenciar em nenhum aspecto. Esse é o caso de **<cc-r>** (*read object - carteira, cupom, bilhete, carta, cf. <sem-r>)* e **<sem-r>** (*read-work - biografia, dissertação, e-mail, ficha cadastral),* que indicam, respectivamente, uma descrição de um objeto de leitura e de um trabalho de leitura. Essas definições sugerem que o que se pretende distinguir é o modo de produção das obras escritas: **<cc-r>** seria relativa àquelas de produção simples, enquanto **<sem-r>**, às de produção complexa. Nesse caso, 'e-mail' e 'ficha cadastral', por requerer produção simples, não deveriam ser exemplos de **<sem-r>.**

Há ainda etiquetas cuja definição se aplica a objetos semanticamente díspares, como **<Adom>** (*Domestic animal or big mammal - terneiro,* leão*/leoa, cachorro),* que, contraditoriamente, trata tanto de animais domésticos quanto de grandes mamíferos. Seria mais conveniente que essa disparidade fosse resolvida com etiquetas mais específicas, que diferenciassem animais domésticos e pequenos mamíferos de animais selvagens ou de grandes mamíferos.

Exemplos do caso (iii) são as etiquetas que, de tão específicas, têm pouca utilidade. Esse é o caso de **<anich>** (*Fish anatomy - few: brânquias, siba)* e **<cc-board>** (*flat long object - few: board, plank, lousa, tabla)*, reconhecidas pelo próprio autor da ferramenta (pela palavra "few" em suas definições) como raramente aplicadas aos itens lexicais de qualquer dos *corpora* investigados.

Caso similar ocorreu com as etiquetas de definições muito abstratas, como **<ac-cat>** (*Category Word - latinismo, número* atômico*)*, corroborando o fato de que as especificações providas para o uso desse elenco não são significativamente esclarecedoras.

O fato de algumas etiquetas serem ontologicamente relacionadas[11] dificultou o processo de revisão dos resultados automáticos, já que muitos itens lexicais podiam ser enquadrados em mais de uma etiqueta (caso (iv)). Isso ocorre, p.ex., com **<fruit>** (*fruit, berry, nut - still mostly marked as <food-c>, abricote, amora, avelã,*

[11] Embora o modelo semântico do PALAVRAS não se baseie em uma ontologia (Bick 2000), é inegável a possibilidade de tratar pelo menos parte delas ontologicamente.

cebola) e **<food-c>** (*countable food - few: ovo, dente de alho, most are <fruit> or <food-c-h> culinary countable food - biscoito, enchido, panetone, pastel).* Certamente, as duas etiquetas são apropriadas para alguns itens lexicais, porém optou-se por utilizar a etiqueta mais específica nesses casos.

Além dos casos acima, ocorrências menos significativas, mas não desprezíveis do ponto de vista da proposta semântica do PALAVRAS, foram elencadas. Verificou-se, dentre elas, que o elenco das 215 etiquetas não foi suficiente para descrever alguns itens lexicais comuns. 'vírus', por exemplo, é etiquetado inadequadamente com **<Acell>** - *Cell-animal (bacteria, blood cells: linfócito)*, pois não é um *animal celular*, mas sim "uma partícula proteica que infecta organismos vivos"[12]. A etiqueta mais próxima a ser atribuída a esse item lexical seria **<cc>** - *concrete countable,* porém, por ser muito genérica, ficou difícil determinar, pelo contexto, sua aplicabilidade. Decidiu-se, assim, manter **<Acell>.** Vale ressaltar que esse foi o único caso de manutenção de etiqueta quando claramente imprópria.

Outras etiquetas são classificadas por Bick como *vazias*, como **<cc-h>** (*artifact, umbrella tag - so far empty category in PALAVRAS)* e parecem se associar a casos não previstos (indicação dada pelo termo *umbrella tag*). No entanto, na ausência de etiquetas adequadas, a escolha pelas ditas *vazias* foi considerada.

Há ainda as marcadas como 'Further proposed categories', para as quais não há definições ou não há exemplos, constituindo-se, assim, em etiquetas subespecificadas. **<spice>** é um caso de ausência completa de descrição; **<top>** (*geographical location)* e **<Bveg>** (v*egetable, espargo, funcho*), de subespecificação.

O uso da etiqueta **<meta>** (*meta noun - tipo, espécie)* também não ficou claro. A referência a *tipo* ou *espécie* sugere a possibilidade de se recorrer a uma relação ontológica. Desse modo, ela poderia ser utilizada para itens lexicais que indicam, por exemplo, classe, gênero ou raça (hiperônimos) de 'equinos' ou 'manga-largas' (hipônimos correspondentes). Decidiu-se por utilizá-la para ocorrências de ambos os tipos, já que nenhuma outra etiqueta do elenco seria apropriada para cobrir esses casos.

Os critérios relatados nesta seção foram adotados mediante a necessidade de se buscar etiquetas adequadas a cada caso, restringindo ao máximo as alterações das anotações originais do *parser*. Ressalta-se ainda que todas as etiquetas constantes do elenco foram utilizadas na pós-edição, razão pela qual confirmamos o alto índice de etiquetas não ocorrentes no corpus (100 ocorrências, ou 47% das etiquetas, não ocorrem no Summ-it).

[12] http://pt.wikipedia.org/wiki/Vírus (Acesso em 25 jun. 2009).

7 Conclusões

Como se demonstrou, o *parser* não dá conta de indicar o conceito semântico adequado para um número significativo de unidades textuais, os quais envolvem, frequentemente, problemas de dependências contextuais e de reconhecimento de entidades nomeadas.

As dificuldades de pós-edição, que implicariam mapeamentos semânticos inadequados dos itens lexicais, foram resolvidas adotando-se vários critérios, dentre os quais o contexto de uso das etiquetas. Evitou-se a opção de adotar etiquetas genéricas quando fosse possível reconhecer alguma mais específica porque essa opção não asseguraria os objetivos do refinamento do VeinSum: ao generalizar etiquetas, a probabilidade de serem indistinguíveis uma unidade textual anafórica e sua antecedente (por suas etiquetas) aumentaria, em vez de diminuir. Assim, embora a etiquetagem semântica de textos de domínios mais genéricos se tenha comprovado menos problemática do que a etiquetagem de textos de domínios mais específicos (que claramente apresentam porcentagem maior de correção), esta opção foi descartada por princípio.

A porcentagem média de correção do corpus (41%) obscurece, certamente, os casos extremos: o texto com menor porcentagem de problemas teve 3% de seus itens lexicais corrigidos; o com maior, aproximadamente 67%. As CCRs referentes a pessoas, as quais, em geral, incluem nomes próprios e profissões, foram as que apresentaram maior índice de acerto.

Considerando-se os vários problemas do *parser* e esses índices de correção da anotação, o corpus pós-editado é um recurso mais rico, pois a atribuição manual de etiquetas foi realizada de forma mais especializada. Resta, assim, sua utilização em tarefas de avaliação ou validação. Particularmente para o modelo de sumarização do VeinSum, será possível validar a revisão das etiquetas verificando se houve melhora da clareza referencial de sumários de outros textos, gerados com base nas heurísticas. Basta compará-los a sumários dos mesmos textos produzidos sem levar em conta as informações semânticas.

De modo geral, claro é que, sem uma reengenharia que envolva critérios semânticos mais robustos do que os atuais, qualquer sistema computacional que dependa da etiquetagem continuará muito vinculado a cada corpus em foco (as heurísticas produzidas, afinal, são dependentes da ocorrência de CCRs que envolvem grupos de etiquetas particulares). Será impossível, no entanto, manter a tarefa de pós-edição manual de resultados semânticos automáticos do PALAVRAS, caso se pretenda que o *parser* semânatico seja um dos módulos de sistemas mais complexos, como o VeinSum. Por outro lado, sua ausência certamente comprometerá a qualidade dos resultados finais do sistema principal.

Assim, seria interessante que houvesse também uma reengenharia do próprio *parser,* para verificar se os problemas aqui detectados de fato podem ser evitados com o refinamento do modelo de etiquetagem. Claramente é necessário, antes, garantir que os problemas de etiquetação apresentados de fato são os causadores da maioria dos problemas de clareza referencial de sumários automáticos gerados pelo VeinSum.

Agradecimentos

Agradecemos a valiosa contribuição dos revisores da revista Linguamática a este artigo. Este trabalho contou com o apoio da FAPESP e da CAPES.

Referências Bibliográficas

Beaugrande, R., W. Dressler. 1981. *Introduction to Text Linguistics.* London, UK, Longman.

Bick, E. 2000. *The Parsing System PALAVRAS: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Arhus, Arhus University.

Carbonel, T. I. 2007. *Estudo e validação de teorias do domínio lingüístico com vistas à melhoria do tratamento de cadeias de correferência em Sumarização Automática.* Dissertação de Mestrado. Departamento de Letras. Agosto. São Carlos, SP, UFSCar.

Chaves, A. R. 2007. *A resolução de anáforas pronominais da língua portuguesa com base no algoritmo de Mitkov.* Dissertação de Mestrado. Departamento de Computação. Agosto. São Carlos, SP, UFSCar: 116p.

Coelho, J. C. B., Muller, V. M., Abreu, S. C., Vieira, R., Rino, L. H. M. 2006. Resolving Nominal Anaphora. *Lecture Notes in Artificial Intelligence 3960,* pp. 160-169. Springer. Berlin, Germany.

Collovini, S., Carbonel, T. I., Fuchs, J. T., Coelho, J. C., Rino, L. H. M., Vieira, R. 2007. Summ-it: Um corpus anotado com informações discursivas visando à sumarização automática. In Violeta Quental, Cláudia Oliveira (eds.), *Proc. of the V Workshop on Information and Human Language Technology* (TIL'2007, CD-ROM). XXVII Congresso da Sociedade Brasileira de Computação (SBC'2007). Rio de Janeiro - RJ.

Cristea, D., Ide, N., Romary, L. 1998. Veins Theory: A Model of Global Discourse Cohesion and Coherence. *Proc. of the Coling/ACL 1998*. Montreal, Canada.

Cristea, D., Postolache, O., Pistol, I. 2005. Summarization through Discourse Structure. *Computational Linguistics and Intelligent Text Processing, 6th International Conference CICLing 2005*. Mexico City, Mexico, Springer LNSC.

Cristea, D., Postolache, O., Puscasu, G., Ghetu, L. 2003. Summarizing Documents Based on Cue-phrases and References. *Proc. of the International Symposium on Reference Resolution and its Apllications to Questions Answering and Summarization.* Veneza, Itália.

Fellbaum, C. 1998. *WordNet: An Electronic Lexical Database.* Cambrigde, Massachussets, The MIT Press

Halliday, M. A. K., Hasan, R. 1976. *Cohesion in English.* London, UK, Longman.

Koch, I. G. V., Travaglia, L. C. 2004. *A coerência textual.* São Paulo, SP, Contexto

Mani, I. 2001. *Automatic Summarization.* Amsterdam, John Benjamin's Publishing Company.

Mann, W. C., Thompson, S. A. 1988. Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. *Text* 8(3): 243-281.

Marcu, D. 2000. *The Theory and Practice of Discourse Parsing and Summarization.* Cambridge, MA, USA, The MIT Press.

Marcuschi, L. A. 1983. *Linguística de texto: como é e o que se faz.* Universidade Federal de Pernambuco. Recife, PE.

Mitkov, R. 1998. Robust pronoun resolution with limited knowledge. *Proc. of the 18th International Conference on Computational Linguistics Conference* (COLING'98/ACL'98). Montreal, Canada.

Mitkov, R. 2002. *Anaphora Resolution.* London, UK, Longman.

Pardo, T. A. S., Nunes, M. G. V. 2002. Segmentação Textual Automática: Uma

Revisão Bibliográfica. *Série de Relatórios Técnicos do Instituto de Ciências Matemáticas e de Computação, no. 185* (NILC-TR-03-02). São Carlos, SP, ICMC, Universidade de São Paulo.

Santos, D. 1990. Lexical gaps and idioms in Machine Translation. *Proc. of the 14th International Conference on Computational Linguistics* (COLING'90), pp. 330-335. H. Karlgren. Helsinque.

Santos, D. 2007. O modelo semântico usado no Primeiro HAREM. In D. Santos, N. Cardoso (eds.). *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*, pp. 43-57, Cap. 4. Linguateca.

Santos, D., Cardoso, N. 2007. Breve introdução ao HAREM. In D. Santos, N. Cardoso (eds.). *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*, Cap. 1. Linguateca.

Sparck-Jones, K. 1999. Automatic Summarizing: factors and directions. In I. Mani, M. Maybury (eds.), *Advances in automatic text summarization,* pp. 1-12. Cambridge, Massachussets: The MIT Press.

Tomazela, E. C. 2010. *O uso de informações semânticas do PALAVRAS: em busca do aprimoramento da seleção de unidades correferentes na Sumarização Automática.* Dissertação de Mestrado. Departamento de Letras. São Carlos, SP, UFSCar. 115p.

Tomazela, E. C., Rino, L. H. M. 2009. O uso de informações semânticas para tratar a informatividade de sumários automáticos com foco na clareza referencial. In Aline Villavicencio (ed.), *Anais do VII Encontro Nacional de Inteligência Artificial (ENIA 2009)*, pp. 799-808. XXIX Congresso da Sociedade Brasileira de Computação. Bento Gonçalves, RS.

Tomazela, E. C., Rino, L. H. M. 2010. *Correção da etiquetagem semântica do Parser PALAVRAS para o Corpus Summ-it.* Série de Relatórios do NILC. NILC–TR-02-10. São Carlos, SP.