# O uso de informações semânticas para tratar a informatividade de sumários automáticos com foco na clareza referencial

**Élen Cátia Tomazela[1], Lucia Helena Machado Rino[2]**

[1]Departamento de Letras
[2]Departamento de Computação
Universidade Federal de São Carlos (UFSCar)
Rodovia Washington Luís, km 235 – SP-310 – 13.565-905 – São Carlos – SP – Brasil
Núcleo Interinstitucional de Lingüística Computacional (http://www.nilc.icmc.usp.br)

etomazela@yahoo.com.br, lucia@dc.ufscar.br

***Abstract.*** *This article presents a refinement proposal of an automatic summarizer by including semantic information in heuristics for text unit selection. The system – VeinSum, is based on three complementary models: Rhetorical Structure Theory, Marcu's Salience Model and Veins Theory. So far, the current heuristics only tackle the problem of avoiding dangling anaphors in the summaries. However, they allow secondary information to be chosen and this may affect its informativeness and even the compression rate. We look into ways of eliminating such superfluous information, so that the summarizer can tackle a more appropriate degree of informativeness. We illustrate cases that may be improved when less information is considered.*

***Resumo.*** *Este artigo apresenta uma proposta de refinamento de um sumarizador automático pela inclusão de informações semânticas às heurísticas de seleção de conteúdo. O sistema – VeinSum – é baseado em três modelos complementares: a Teoria RST, o Modelo de Saliência de Marcu e a Teoria das Veias. As heurísticas atuais tentam evitar que anáforas sejam incluídas sem seus respectivos antecedentes. Porém, elas permitem a escolha de informações secundárias que prejudicam a informatividade e também a taxa de compressão do sumário. Tentamos eliminar informações supérfluas para que o sumarizador assegure uma melhor informatividade. Ilustramos casos que podem ser melhorados quando menos informação é considerada.*

## 1. Introdução

A resolução de correferência é um tópico que requer atenção especial em várias áreas de Processamento das Línguas Naturais, tais como Extração de Informação, Tradução e Sumarização Automáticas (SA). É possível, nos dias de hoje, encontrar sumarizadores com desempenho satisfatório quando se trata da fidelidade de conteúdo. Já quando se trata de textualidade, o desempenho deixa a desejar, levando a dificuldades de compreensão.

Os sistemas que apresentam processamento lingüístico profundo deveriam garantir que a informação selecionada para compor um sumário fosse organizada e coerentemente reproduzida de forma condensada a partir de uma representação lingüística do conteúdo do texto-fonte. Entretanto, devido à complexidade de se identificar adequadamente a interdependência entre as unidades textuais, a coerência dos sumários é frequentemente prejudicada. De nosso especial interesse[1], visando melhorá-la, é a preservação da *clareza referencial* nos sumários, definida como a propriedade que um texto tem de permitir ao leitor identificar a quem ou a que um determinado pronome ou sintagma nominal está se referindo[2]. Caso os sumários produzidos não apresentem essa propriedade, o entendimento de seu conteúdo, e assim sua qualidade e utilidade, podem ser comprometidas.

Esse já é o foco do sumarizador automático considerado, o VeinSum [Carbonel et al., 2007], o qual se baseia em duas teorias de organização do discurso: a Teoria das Veias, ou VT [Cristea et al., 1998] e a RST [Mann & Thompson, 1988] . A VT se baseia numa árvore RST para determinar um conjunto de unidades de discurso que possa conter possíveis antecedentes de uma expressão anafórica, denominado *domínio de acessibilidade referencial*, ou *acc,* o qual deve também ser incluído no sumário caso a unidade textual que contém uma anáfora o for. Para determinar as unidades que farão parte de um sumário, o VeinSum deve obedecer simultaneamente à VT e ao Modelo de Saliência [Marcu, 1997]. Este torna obrigatória a inclusão de unidades textuais por sua ordem de classificação de saliência e não por sua interdependência referencial. Por esse motivo, esse modelo é associado à VT: a cada unidade saliente incluída, acrescenta-se no sumário seu *acc* inteiro, para que não haja quebras de clareza referencial. O problema, aqui, é que a VT determina o *acc*, sobretudo, topologicamente, isto é, ela não faz uso explícito de conhecimento semântico para descobrir quais unidades são correferentes. Assim, o cálculo do *acc* não é preciso e ele pode conter unidades textuais que nada têm a ver com o contexto referencial de uma anáfora, levando à inclusão de trechos de importância secundária. Estes, por sua vez, podem depreciar a informatividade dos sumários, se comparados aos seus textos-fonte. Esse problema pode se agravar ainda mais sob a restrição de compressão: o risco de violar essa taxa é alto devido à necessidade de inclusão do *acc* de cada unidade de informação. Neste caso, o VeinSum despreza unidades mais salientes em prol de unidades cujos *accs* favoreçam tal taxa.

Propomos, assim, diminuir os *accs* de EDUs que contenham expressões anafóricas, sem, contudo, interferir na classificação de saliência original do sistema. Para isso, utilizamos informações semânticas provindas do *parser* PALAVRAS [Bick, 2000] e, caso essas sejam insuficientes, de outros recursos lingüísticos, como a WordNet [Fellbaum, 1998].

Este artigo está organizado da seguinte forma: na seção 2, descrevemos os modelos fundamentais do VeinSum e o seu modo de funcionamento. Na seção 3. , apresentamos uma análise detalhada de como as informações semânticas podem ser úteis para o nosso propósito e, finalmente, na seção 4, são apresentadas as considerações finais.

---

[1] Proposta inicial do Projeto ProCaCoSA, Proc. CNPq Nro. 503766/2005-4.

[2] Definição utilizada na DUC2005 (http://duc.nist.gov/duc2005/).

## 2. VeinSum: Um sumarizador automático de textos em português

Detalhamos, nesta seção, os modelos fundamentais utilizados pelo VeinSum e ilustramos detalhadamente o seu funcionamento. Além disso, apresentamos também suas limitações sob a ótica de construção de sumários que assegurem clareza referencial e um nível adequado de informatividade.

### 2.1 Os Modelos Fundamentais do VeinSum

O VeinSum[3] foi projetado para sumarizar estruturas RST e não textos em língua natural diretamente. Ou seja, supõe-se que haja um analisador discursivo prévio que, ao receber um texto, produza sua árvore RST, a qual é construída gradualmente pela junção de suas unidades elementares, ou EDUs (do inglês, *Elementary Discourse Units*), através de relações retóricas, ou relações RST. A cada EDU é atribuído o papel de núcleo (N) ou satélite (S), dependendo do grau de importância que tal unidade expressa no texto. Relações RST podem ser de dois tipos: *mononucleares* (entre um N e um S) ou *multinucleares* (entre vários Ns). São as relações mononucleares que permitem delinear EDUs supérfluas e, assim, candidatas à exclusão de um sumário[4]. Essa forma privilegia o reconhecimento de segmentos relevantes já na representação linguística do texto-fonte, muito embora as folhas da árvore ainda sejam seus próprios segmentos textuais elementares. A RST já foi explorada para a SA por diversos autores: (p. ex., [Marcu, 1997], [Marcu, 1999], [Marcu, 2000], [O'Donnell, 1997], [Ono et al., 1994]) devido à nuclearidade dos segmentos textuais representados, que, em princípio, delinearia uma estratégia simples de excluir Ss e manter os Ns [Sparck Jones, 1993]. Um N, nesse caso, expressaria informação mais saliente, quando comparado ao seu S, cuja exclusão não prejudicaria o entendimento da mensagem. O problema de se construir sumarizadores baseados somente na RST é que, além de não se poder excluir indiscriminadamente os Ss, fica impossível tratar o fenômeno do encadeamento referencial, já que as relações RST conectam unidades discursivas inter-oracionais e não contempla qualquer informação intraoracional. De qualquer modo, agregada à RST, ela é usada para a SA de textos, seguindo a proposta de Cristea et al. [1998].

Com a RST e a VT, é possível delinear os contextos referenciais, mas não a escolha das EDUs para compor um sumário, função exercida pelo Modelo de Saliência, cuja premissa é que as EDUs que estão mais próximas da raiz da árvore são mais importantes do que aquelas que se encontram em níveis mais profundos. Neste caso, elas têm peso maior para compor um sumário e o cômputo desse peso, ou saliência, se baseia tanto na nuclearidade quanto na profundidade das EDUs na estrutura RST. Aplicando-se o algoritmo de Marcu correspondente, obtemos uma classificação das EDUs e a usamos para determinar a ordem de escolha para produzir o sumário. Por fim, agregando os três modelos propostos, contemplamos a clareza referencial ao usar o *acc* como o conjunto de EDUs correferenciais a uma EDU que contém uma anáfora (somente as EDUs antepostas a ela são consideradas). Segundo Cristea et al., o *acc* indica o contexto mínimo necessário em que é possível recuperar o antecedente de uma

[3] Descrição detalhada em Carbonel [2007].

[4] Afinal, é impossível distinguir diferentes graus de importância para EDUs multinucleares, por isso, a inclusão de todas elas deveria ser considerada.

anáfora e, assim, assegurar que a mensagem expressa no texto seja clara para o leitor. Porém, ao levar em conta somente a posição que as EDUs ocupam em uma árvore RST e sua nuclearidade, verificamos que os *accs* determinados pelo VeinSum ainda podem conter EDUs que nada têm a ver com o contexto referencial da anáfora. Ou seja, os modelos atuais do sistema são insuficientes para a resolução desse problema, o qual remete a problemas de informatividade. Daí a proposta de considerarmos a inclusão de processamento semântico no VeinSum, mantendo os mesmos modelos teóricos para assegurar a clareza referencial, mas visando melhorar a informatividade dos sumários, como discutimos na Seção 3.

## 2.2 Funcionamento do VeinSum

A arquitetura do VeinSum exibida na Figura 1 (Carbonel, [2007, p.119]) destaca seus principais módulos. Seu funcionamento é ilustrado aqui para o excerto designado Texto1 (com EDUs seguindo a numeração do texto-fonte), já etiquetado semanticamente pelo PALAVRAS (as etiquetas estão indicadas entre <...> e seu uso é descrito na Seção 3)[5]. Estamos interessados, particularmente, na EDU7, a qual contém uma descrição definida (DD) anafórica (em negrito, com antecedente sublinhado).

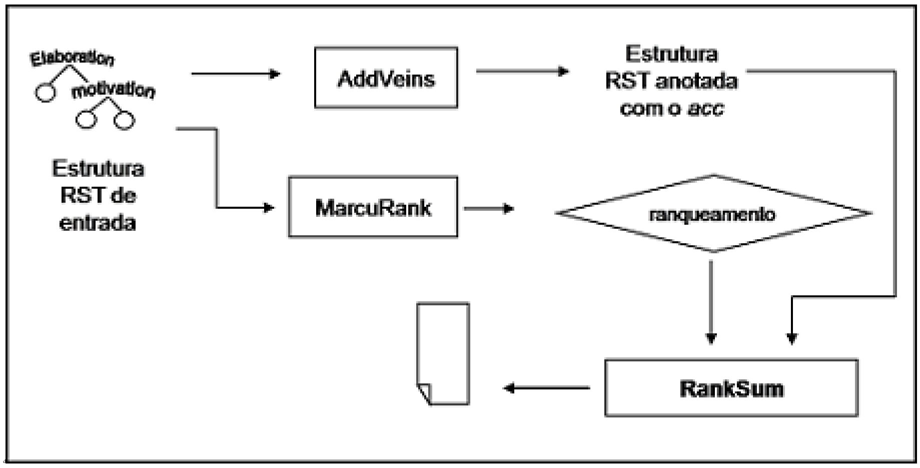


**Figura 1. Arquitetura do VeinSum**

Numa entrevista coletiva conduzida ontem à noite **<temp>,** os gerentes **<Hprof>** da Nasa **<org>** deram o veredicto **<act-s>**. [1] O Discovery **<Vair>** precisa de reparos **<f-c>** [2] antes de voltar para casa. [3] O trabalho **<act-d>** será conduzido amanhã pelo <u>astronauta</u> **<Hprof>** <u>americano</u> **<Hnat>** <u>Stephen Robinson</u> **<hum>**, durante a terceira caminhada **<activity>** espacial prevista pela missão **<occ>**. [4] Sua tarefa **<act-d>** consistirá em cortar arestas **<percep-w>** do material **<mat>** usado para preencher o vão **<part-build>** entre as telhas térmicas **<Lcover>** da barriga do ônibus **<part-build>.** [5] Para chegar lá, [6] **o tripulante <Hprof>** será preso à ponta do braço robótico **<part-build>** da ISS. [7]

**Texto1. Trecho do texto CIENCIA_2005_28754**

[5] Todos os excertos ilustrados neste artigo são extraídos de textos do *Corpus* Summ-it [Collovini et al., 2007].

Para esse segmento, a árvore RST de entrada para o VeinSum contém a subárvore exibida na Figura 2 e o sumário automático resultante contém o excerto denominado Texto2 correspondente.

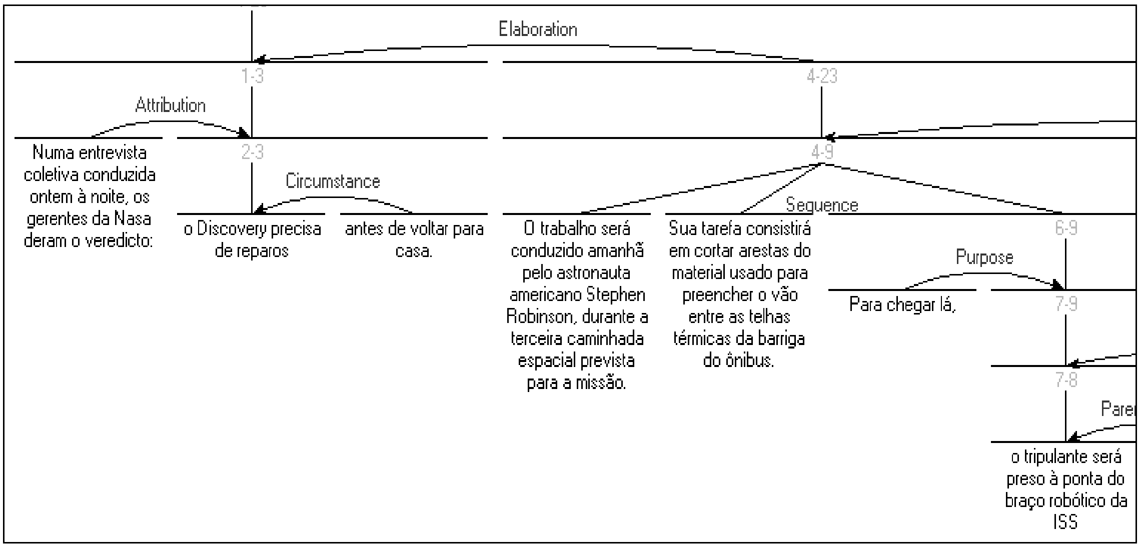


**Figura 2. Estrutura RST do Texto1**

O Discovery precisa de reparos.[2] O trabalho será conduzido amanhã pelo <u>astronauta Stephen Robinson</u>, durante a terceira caminhada espacial prevista pela missão.[4] Sua tarefa consistirá em cortar arestas do material usado para preencher o vão entre as telhas térmicas da barriga do ônibus.[5] Para chegar lá[6], **o tripulante** será preso à ponta do braço robótico da ISS.[7]

**Texto2. Excerto do sumário do VeinSum produzido para o Texto1**

Seguindo a classificação de saliência de todas as EDUs do Texto1, realizada pelo módulo *MarcuRank,* o VeinSum pode ter selecionado outras EDUs para o sumário, antes de chegar à EDU7. Independentemente de quais sejam, o que podemos notar é que a inclusão dessa EDU, com seu *acc*, não viola a taxa de compressão, caso contrário, ela não estaria no Texto2. Vejamos os passos do VeinSum para sua inclusão:

1) Recupera seu *acc*, calculado pelo módulo *AddVeins*: *acc*(7) = {2,4,5,6,7}.
2) Inclui, além da EDU7, as demais EDUs desse conjunto, ante a suposição de que alguma delas conterá o antecedente da anáfora. Note-se que a ordem em que elas aparecerão no sumário é a mesma do texto-fonte.
3) Calcula o tamanho do sumário até o momento. Se ele não viola a taxa de compressão, busca a próxima EDU saliente como candidata à inclusão. Caso contrário, descarta a EDU7 (e conseqüentemente seu *acc*) e elege a próxima na classificação de saliência.

Esses mesmos passos de raciocínio são repetidos até que o sumário atinja a taxa de compressão. Para a construção do sumário completo do Texto2 não houve necessidade de nenhum descarte, como comprova a escolha de todas as unidades mais salientes indicadas pelo Modelo de Saliência, conforme ordem gerada automaticamente: 2>40>34,41,44,47,49>4,5,7>.....> ....6 ...>...

Na verdade, a espinha dorsal para a decisão das EDUs incluídas é dada por *spine*="[2,40,34,41,44,47,49,4,5,7]". A inclusão das demais EDUs nesse sumário (algumas aparentes no Texto2) se deve ao passo (2) acima, tanto para a EDU7 quanto para outras EDUs também selecionadas a partir dessa classificação. Nota-se que a EDU6 não seria incluída levando em conta somente a *spine*, mas ela faz parte do *acc*(7).

Diretamente podemos ver que no Texto2 há EDUs do *acc* que poderiam ter sido descartadas sem prejuízo da clareza referencial em relação à EDU7. Porém, como já dissemos, o VeinSum não trata dessa questão, daí nossa proposta de uso de informações semânticas para aprimorá-lo, como apresentamos na próxima seção.

## 3. O uso de informações semânticas para tratar a informatividade de sumários automáticos com foco na clareza referencial

Nosso objetivo é o de descartar do *acc* de cada EDU apontada pelo Modelo de Saliência as que nada têm a ver com o seu contexto referencial e, caso essa EDU contenha uma anáfora, assegurar sua clareza referencial. Ao eliminar informações secundárias do *acc*, a classificação de saliência do sistema será mais fielmente obedecida e, assim, o sumário terá um melhor grau de informatividade, ou seja, seu conteúdo refletirá melhor o conteúdo principal do texto-fonte. Para isso, propomos duas abordagens metodológicas complementares: o reconhecimento de entidades correferentes por sua similaridade prototípica, através das etiquetas providas pelo *parser* PALAVRAS e o uso da WordNet quando o reconhecimento não for possível através da primeira proposta. Nossa hipótese principal é que todos os membros de uma CCR recebam ou as mesmas etiquetas semânticas ou etiquetas similares, já que todas as menções se referem à mesma entidade. São considerados somente os substantivos de sintagmas nominais, pois outros tipos de anáfora (como a pronominal) não possuem etiquetas semânticas próprias.

A similaridade prototípica, como definida por Bick, consiste na propriedade de itens lexicais distintos compartilharem conjuntos de traços semânticos pela similaridade de seu significado (e não por sua coincidência semântica). A medida de similaridade é proporcional ao número de traços semânticos que eles compartilham: Bick supõe que, quanto maior esse número, mais similares eles serão. Nessa abordagem, itens similares podem não ser sinônimos, mas partilhar o mesmo campo semântico. A Figura 3 ilustra, em cada linha da tabela parcial, protótipos [Bick, 2000, p.307] que compartilham os mesmos traços semânticos, estes representados pelas letras do cabeçalho da tabela. Ela é usada para demonstrar o uso desses protótipos e verificar que as informações semânticas providas pelo *parser* de fato são úteis na identificação de expressões correferentes.

Voltando ao excerto Texto1, pode-se verificar que as unidades correferentes à DD **o tripulante,** que recebe etiqueta **<H>,** são <u>Stephen Robinson</u> e <u>o astronauta</u>, sintagmas nominais etiquetados respectivamente como **<hum>** e **<Hprof>.** A primeira linha da tabela indica que essas três etiquetas (marcadas em caixa) compartilham os mesmos traços semânticos e fazem parte do mesmo protótipo. Isso indica que esse protótipo pode ser suficiente para delimitar possíveis antecedentes da EDU7.

E = entities (±CONCRETE) V = ±VERBAL
C = ±CONTROL P = ±PERFECTIVE
I = ±MOVING S = ±MEASURING
J = ±MOVABLE D = ±PARTITIVE
A = ±ANIMATE (living) X = ±HUMAN-EXPRESSION (allowing human modifier-ADJ)
H = ±HUMAN ENTITY F = feature (±ADJECTIVAL)
M = ±MASS L = ±LOCATION
N = number (±COUNTABLE) T = ±TEMPORAL

| E | C | I | J | A | H | M | N | V | P | S | D | X | F | L | T | Cluster |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| + | | '+ | + | '+ | '+ | . | '+ | | | . | | + | . | . | . | H, Hprof, Hnat, Hmyth, Hfam, Htit, i, Hbio, Hsick, Hattr, *hum |
| + | | '+ | + | '+ | '+ | . | . | | | . | | + | . | . | . | HH, Hhparty, *party, *media |
| + | | | . | . | '+ | . | . | | | . | | + | . | '+ | . | inst, *inst |
| + | | | . | . | '+ | . | '+ | | | . | | + | . | '+ | . | Lciv, *civ |
| + | | '+ | + | '+ | . | . | '+ | | | . | | . | . | . | . | A, Azo, Aorn, Aent, Aich, Amyth, Acell, Adom |

**Figura 3. Traços semânticos correspondentes às etiquetas**

Com nossa proposta de escolher do *acc* somente as EDUs que possuem etiquetas similares, as EDUs 5 e 6 seriam desconsideradas, pois elas não contêm substantivos cujas etiquetas façam parte do mesmo protótipo de **<H>**, muito embora o VeinSum acuse-as como pertencentes ao *acc*(7). Teríamos, então o seguinte sumário:

> O Discovery precisa de reparos.[2] O trabalho será conduzido amanhã pelo astronauta Stephen Robinson, durante a terceira caminhada espacial prevista pela missão.[4] **O tripulante** será preso à ponta do braço robótico da ISS.[7]

**Texto3. Exemplo de sumário manual do Texto1**

Embora menos informações sejam veiculadas, podemos perceber que as informações excluídas podem ser consideradas de importância secundária, não prejudicando a veiculação da CCR restante, ou seja, é possível assegurar a clareza referencial de modo que restem somente as EDUs mais informativas para essa CCR. Repare ainda que a EDU2 também poderia ser excluída pelo mesmo motivo. No entanto, ela já faz parte do sumário, pois ao incluir a EDU4, o *acc*(4) =[2,4] já foi considerado, garantindo que a anáfora **o trabalho** também tenha o seu antecedente explícito.

Também encontramos casos cujas etiquetas semânticas não são suficientes para determinar um subconjunto ideal do *acc*, como no excerto denominado Texto4. O *acc*(26) = [3,6,8,24,25,26] também poderia ser diminuído caso o foco fosse na EDU26, que contém a anáfora **o clone**. As EDUs correferentes a ela, nesse caso, são as EDUs 24 e 25, resultando nas etiquetas <Adom>, <Azo> e <A>, esta última sendo a etiqueta da anáfora em foco. No entanto, a EDU3, que também faz parte do *acc*(26), não seria excluída porque também é etiquetada com <Azo>. Ou seja, uma heurística que exclua EDUs com etiquetas em protótipos distintos do protótipo do termo anafórico já não serviria aqui (todas elas estão no protótipo indicado na 5ª. linha da tabela da Figura 3).

> Duas vacas**<Azo>**deram cria ontem em Iowa.[3]Os dois nascimentos[6]marcam o início de uma nova fase para um projeto que já atraía interesse.[8]A primeira tentativa de trazer um membro do Frozen Zoo de volta do mundo dos animais perdidos foi com um gauro **<Adom>**,outra espécie rara de gado**<Adom>.**[24]Uma gravidez acabou levando ao nascimento de um animal**<Azo>** em 2001.[25] **O clone <A>** morreu dois dias depois. [26]

**Texto4. Trecho do texto CIENCIA_2003_24212**

O excerto Texto5 mostra que parte do objetivo de assegurar a clareza referencial e melhorar a informatividade é atingido. Entretanto, a EDU3 trata de um tópico distinto do indicado pela CCR em foco e, se incluída no sumário, danificará a informatividade, pois não pertence ao mesmo contexto da CCR. Uma forma de descartá-la seria utilizando também as informações de gênero e número.

> Duas vacas deram cria em Iowa. A primeira tentativa de trazer um membro do Frozen Zoo de volta do mundo dos animais perdidos foi com um gauro, outra espécie rara de gado. Uma gravidez acabou levando ao nascimento de um animal, em 2001. O clone morreu dois dias depois.

**Texto5. Exemplo de sumário manual do Texto4**

Há casos, ainda, em que nem informações semânticas ou morfossintáticas são suficientes. Nesses casos, pode-se fazer uso dos conceitos da WordNet a fim de trabalhar com o relacionamento hierárquico dos *synsets* e encontrar uma posição comum na hierarquia em que os itens lexicais possam estar relacionados. Em outro texto do *Corpus* Summ-it, por exemplo, encontramos uma CCR com os itens lexicais **o texto <sem-r>** e a declaração de Helsinque **<occ>.** Pela similaridade de etiquetas o reconhecimento automático de que esses itens são correferentes fica difícil, pois, por sua definição, elas não poderiam fazer parte de um mesmo protótipo. Neste caso a busca na WordNet pode ser útil, pois é possível chegar ao *synset* hiperônimo de suas respectivas traduções *text* e *declaration*, o qual é *communication*. O próximo passo é, então, buscar uma etiqueta representativa desse *synset* no conjunto provido pelo *parser* e usá-la como meio de enquadrar os itens em uma única CCR.

Caso não seja possível reconhecer sequências de itens lexicais correferentes e, assim, minimizar um *acc*, entendemos que não é possível resolver o problema em foco. No entanto, ainda não comprovamos essa proposta de uso de informações semânticas para tratar a informatividade dos sumários automáticos. Nosso próximo passo será enriquecer as heurísticas do sumarizador para contemplar os casos aqui ilustrados.

## 4. Considerações Finais

O trabalho aqui relatado dá continuidade ao Projeto ProCaCoSa, o qual teve por objetivo construir sumários que não contivessem anáforas sem suas CCRs completas e resultou justamente no VeinSum. Esse sumarizador automático, embora ainda em estágio de protótipo, foi o primeiro voltado ao processamento de estruturas RST para a SA de textos em português. Por essa razão, embora o sistema apresente ainda muitos problemas, ele é nossa perspectiva de continuidade de uma pesquisa de cinco anos. A utilização do VeinSum não é trivial, sobretudo porque ele não admite textos reais como entrada, mas suas estruturas RST. Embora a perspectiva do ProCaCoSA fosse ter um analisador discursivo já acoplado ao VeinSum, essa situação ainda não se configura devido a problemas de desempenho dos diversos módulos interdependentes e interagentes. Além disso, por seguir a abordagem profunda, o esforço humano de modelagem, triagem e até de reengenharia é muito grande. Por esses motivos, até o momento avaliações amplas não são possíveis: delas depende não só a construção de *corpora* de referência (também manual), como a construção manual de dados de entrada (estruturas RST), para testar o módulo de SA isoladamente.

As heurísticas em perspectiva estão em construção, porém, devido ao uso de módulos inteiramente automáticos para a manipulação do conhecimento, como é o caso da etiquetagem semântica pelo PALAVRAS, deparamo-nos com inúmeros outros problemas de cuja resolução depende a continuidade da proposta. Por exemplo, decidimos corrigir a etiquetagem atual do *Corpus* Summ-it (como fizemos com as etiquetas ilustradas antes) porque observamos que havia um alto índice de inadequações de etiquetas. Essa correção é feita manualmente por uma especialista lingüista. Durante essa revisão, também descobrimos vários casos de etiquetagem morfossintática imprecisa, o que certamente influirá nas decisões de escolha de EDUs que temos como foco neste trabalho. Como este é um trabalho básico de exploração da semântica lexical, entendemos que esses problemas não podem ser ignorados.

Outras dificuldades que também antevemos é que, ao diminuir um *acc*, pode acontecer de o conjunto resultante de EDUs ser insuficiente, danificando a clareza referencial. Esta é uma questão de pesquisa que só poderá ser explorada a partir da proposta de refinamento das heurísticas de seleção. O que notamos, no entanto, é que, com o modelo atual do VeinSum, o qual despreza unidades mais salientes buscando respeitar a taxa de compressão, aumenta-se a probabilidade de o sumário ter um nível de informatividade que não corresponde ao nível principal pretendido com o texto-fonte. Em caso extremo, ele deixará de transmitir a idéia central do texto, pois poderá contemplar somente informações periféricas. Em outras palavras, o equilíbrio entre os critérios de informatividade, clareza referencial e compressão é, em si, um problema desafiador em nossa proposta.

Citamos, abaixo, alguns trabalhos relacionados ao nosso que perseguem o objetivo de não danificar a clareza referencial: o CorrefSum [Gonçalves, 2008] é um sistema que prevê a pós-edição de sumários extrativos pela substituição de expressões anafóricas pelos componentes de sua CCR. Porém, a substituição indiscriminada pode prejudicar a informatividade e mesmo a taxa de compressão. Já Azzam et al. [1999] utilizam a marcação de CCRs para construir sumários que focam em uma única entidade. Tal sumário é construído baseado em uma variedade de critérios que selecionam, dentre todas as CCRs que compõem o texto-fonte, a que melhor representa o seu conteúdo. A idéia principal, nesse caso, é que um texto é em grande parte construído em torno de uma entidade central, considerado o foco ou tópico do discurso. Cristea et al [2005] também utilizam estruturas parecidas com a RST e baseiam-se na nuclearidade e na topologia das árvores para produzir sumários que focam em uma única entidade mencionada no texto. A VT é aqui utilizada para guiar a construção das árvores e no processo de SA em si.

Nossos próximos passos visam à construção das heurísticas baseadas em informações semânticas e a elaboração de avaliações que simulem o desempenho do VeinSum.

## Agradecimentos

Este trabalho tem o apoio da FAPESP, CNPq e CAPES.

## Referências Bibliográficas

Azzam, S., Humphreys, K. and Gaizauskas, R. (1999) "Using correference chains for text summarization", Workshop on Conference and its Applications, pp. 77-84, Baltimore.

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.

Carbonel, T. I., Pelizzoni, J. M. and Rino, L. H. M. (2007) "VEINSUM: Um Modelo de Sumarização Automática de Textos Baseado em Estruturas Retóricas", CoPG - Congresso de Pós-Graduação da USFCar, São Carlos - SP.

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

Cristea, D., Ide, N. and Romary, L. (1998) "Veins Theory: A Model of Global Discourse Cohesion and Coherence", In: Proc. of the Coling/ACL 1998, pp. 281-285.

Cristea, D., Postolache, O. and Pistol, I. (2005) "Summarization through Discourse Structure", In: Computational Linguistics and Intelligent Text Processing, 6th International Conference CICLing 2005, Edited by A. Gelbukh, pp. 632-644, Mexico City, Mexico, Springer LNSC.

Fellbaum, C. D. (1998), WordNet: an electronic lexical database, Cambridge, The MIT Press.

Gonçalves, P. N. (2008) "CorrefSum: Revisão de Coesão Referencial em Sumários Extrativos", Dissertação de Mestrado, Departamento de Computação, Agosto, pp. 129. São Leopoldo, RS, Universidade do Vale do Rio dos Sinos.

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

Marcu, D. (1997) "The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts", Computer Science, pp. 351, Toronto, Canada, University of Toronto.

Marcu, D. (1999) "Discourse trees are good indicators of importance in text", In: Advances in Automatic Text Summarization, Edited by I. Mani and M. Maybury, pp. 123-136, The MIT Press.

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

O'Donnell, M. (1997) "Variable-length on-line document generation", In: Proceedings of the 6th European Workshop on Natural Language Generation, pp. 82-91, Gerhard-Mercator University, Duisburg, Germany.

Ono, K., Sumita, K. and Miike, S. (1994) "Abstract generation based on rhetorical structure extraction", In: Proceedings of 15th International Conference on Computational Linguistics (COLING'94), pp. 344-348, Kyoto, Japan.

Sparck Jones, K. (1993) "What might be in a summary?", In: Information Retrieval 93, Edited by G. Knorz, J. Krause and C. Womser-Hacker, pp. 9-26, Konstanz, Universitatsverlag Konstanz.