Validação Preliminar da Teoria das Veias para o Português e Lições Aprendidas

Thiago Ianez Carbonel[1], Jorge Marques Pelizzoni[2], Lucia Helena Machado Rino[3]

[1]Departamento de Letras – UFSCAR; [2]Instituto de Ciências Matemáticas e de Computação – USP; [3]Departamento de Computação – UFSCAR

Núcleo Interinstitucional de Lingüística Computacional (NILC)

thiagocarbonel@gmail.com, jpeliz@icmc.usp.br, lucia@dc.ufscar.br

Abstract. *In this paper we report on the first validation effort for Portuguese for Veins Theory's Conjecture 1 (C1), which constrains anaphora resolution given the rhetoric structure of texts and whose applicability to Automatic Summarization interests us. Our experiment is preliminary in that only definite noun phrases are considered. Thereby we were able to improve our methodology and observe a rather lower precision than reported elsewhere for other languages. As a methodological novelty, we put forth the* Non-Trivial Precision*, a more realistic estimator of C1's predictive power.*

Resumo. *Neste artigo, reportamos o primeiro esforço de validação da Conjectura 1 (C1) da Teoria das Veias para o português, a qual restringe a resolução anafórica em função da estrutura retórico-discursiva de um texto e cuja aplicabilidade à Sumarização Automática nos interessa. Trata-se de um experimento preliminar, considerando apenas sintagmas nominais definidos. Por meio dele, pudemos aperfeiçoar nossa metodologia e observar uma precisão bem menor que a reportada anteriormente para outras línguas. Como inovação metodológica, formulamos a* Precisão Não-Trivial*, um estimador mais realista para o poder preditivo da C1.*

1. Introdução

Dentro da pesquisa relacionada à resolução anafórica, Cristea et al. (1998) formularam a Teoria das Veias (*VT*, do inglês *Veins Theory*), que relaciona a estrutura retórico-discursiva (arbórea) de um texto e as cadeias de co-referência que nele se desenvolvem. Resumidamente, dado um elemento anafórico qualquer dentro de um texto, a VT postula restrições sobre o seu *domínio de acessibilidade referencial*, que é definido como o contexto de ocorrência de elementos que lhe sejam co-referentes e, em especial, aquele que primeiro menciona o seu referente no texto. Em específico, a VT afirma que a procura por referentes no texto não precisa levar em consideração todo o texto que precede uma anáfora, mas apenas uma porção bem menor deste, a qual é determinada em função (i) da árvore retórico-discursiva do texto e (ii) do local onde exatamente a anáfora se situa nessa

estrutura. A essa afirmação deu-se o nome de *Conjectura 1*, que será referenciada neste artigo por C1.

Assumindo-se a validade da VT, um corolário interessante diz que, dada uma unidade discursiva do texto (uma sentença ou mesmo uma oração) qualquer, pode-se determinar uma porção relativamente pequena do texto antecedente que sabemos conter todos os requisitos para que essa unidade "faça sentido", na medida em que não haverá perda de referência nesse excerto – dito *texto mínimo* – composto apenas dessa porção e da unidade em questão. Esse é um resultado especialmente bem-vindo para a Sumarização Automática, extrativa, que obtém sumários exclusivamente por meio de operações de deleção sobre o texto original. É que, caso um sumarizador extrativo seja bem-sucedido ao eleger os fragmentos mais importantes de um texto, a VT oferece subsídios para complementá-los com outros fragmentos que, apesar de menos centrais, sejam provavelmente necessários para a enunciação daqueles. Essa abordagem já se provou promissora por Seno (2005) e Seno & Rino (2005) em seu sistema RHeSumaRST, que utiliza o cálculo de veias em conjunto com um modelo de saliência (Marcu, 1997; 1999) e heurísticas de poda visando, principalmente, a manutenção dos elos referenciais em sumários produzidos automaticamente.

Cristea et al. fizeram experimentos para verificar a validade da VT para as línguas inglesa, francesa e romena, demonstrando ter a teoria um poder preditivo significativo – na verdade, quase absoluto. Por sua vez, o trabalho de Seno e Rino cobre a única aplicação já feita da teoria para o português e, apesar dos resultados promissores obtidos, não incluem uma validação para esta língua nos moldes de Cristea et al. Em outras palavras, o objetivo de Seno e Rino foi apenas demonstrar que a teoria pode auxiliar na geração de sumários mais coerentes, e não produzir resultados comparáveis aos de Cristea et al.

O objetivo primeiro deste trabalho é exatamente preencher essa lacuna, ainda que parcialmente. Apresentamos aqui os resultados de um esforço preliminar de validação da *C1*, obtidos para um córpus de textos de divulgação científica anotado manualmente quanto à (i) co-referência e (ii) estruturação retórico-discursiva. Restringimo-nos à C1 por ser a mais pertinente à questão da resolução anafórica e, portanto, à manutenção da coerência referencial em sumários, nosso objeto de pesquisa original. Nosso experimento sofre de uma limitação importante: no momento, dispomos de anotação de co-referência completa apenas para os sintagmas nominais definidos (isto é, aqueles iniciados por artigo definido) de nosso córpus. Tomamos o cuidado, entretanto, de aí incluir todos os antecedentes, mesmo que de outras categorias, desses elementos. Portanto, estritamente falando, verificamos a validade da C1 apenas para essa (importante) categoria de expressões referenciais.

Como objetivo secundário, relatamos as lições aprendidas no processo de depurar nossos resultados. Em outras palavras, sempre que detectávamos um caso não coberto pela C1, cumpria verificar se a real causa da falha não seria um eventual erro de anotação. Assim, pudemos não só melhor abordar os casos realmente não cobertos pela C1, mas também traçar críticas e apontamentos interessantes sobre nosso modelo de anotação.

Importante ressaltar que o trabalho apresentado neste artigo não visa a uma comparação entre os experimentos realizados por Cristea et al. e o nosso, mas sim reportar um experimento de validação da VT para o português, no qual explicitamos diferenças metodológicas e apresentamos dados mais realistas obtidos exatamente em função de uma abordagem experimental voltada apenas aos casos não triviais, compreendidos estes como os casos realmente significativos no nosso contexto.

Todo o trabalho aqui reportado foi desenvolvido no âmbito do Projeto ProCaCoSA – PROcessamento de Cadeias de CO-referência para a Sumarização Automática de Textos em Português[1] – que visa analisar e diagnosticar problemas causados à Sumarização Automática pela ocorrência de cadeias de co-referência não resolvidas durante a seleção e estruturação do conteúdo de sumários, situação que caracteriza a perda da referência por conta de quebras de elos referenciais nos sumários, antes assegurados no texto-fonte. Para investigar esse tipo de déficit de textualidade e instanciar o uso da VT, utilizamos o modelo RST (Mann & Thompson, 1987) de anotação retórico-discursiva, quer na fase de estruturação dos textos completos, quer na fase de seleção e estruturação de seus sumários.

Este artigo prossegue da seguinte forma: na Seção 2, introduzimos resumidamente a RST, a VT e suas inter-relações. Descrevemos nossa metodologia de investigação, bem como o experimento propriamente dito e seus resultados, na Seção 3. A seguir, na Seção 4, apresentamos e discutimos as conseqüências do processo de depuração dos resultados para nossa prática de anotação. Por fim, concluímos e mencionamos trabalhos futuros Seção 5.

2. Teoria das Veias e Teoria de Estrutura Retórica

A Teoria das Veias (VT) generaliza o conceito de coerência local proposto por Grosz et al. (1995) e parte de uma hipótese de análise retórico-discursiva nos moldes da Teoria de Estrutura Retórica (RST), desenvolvida por Mann & Thompson (1987). Portanto, cumpre introduzir primeiro a RST.

A RST fundamenta-se no princípio de que um texto tem uma estrutura retórico-discursiva – estritamente arbórea e, portanto, recursiva – que subjaz a estrutura superficial. Ela supõe que, dada essa estrutura, é possível recuperar o objetivo comunicativo que o escritor do texto pretendeu atingir ao escrevê-lo. O texto – ou discurso, na acepção aqui adotada – pode ser segmentado em unidades mínimas de significado, denominadas EDUs – do inglês *Elementary Discourse Units* (Marcu, 1997) – que, necessariamente, mantêm relação entre si na construção textual. Em uma árvore retórica, ou árvore RST, as EDUs constituem suas folhas, enquanto cada relação corresponde a seus nós intermediários, cujos rótulos indicam o tipo da relação retórica e cujas subárvores são os relacionandos. Trata-se de árvores ordenadas, sendo que a leitura das folhas (EDUs) da esquerda para a direita resulta na reconstituição do texto original.

A teoria prevê um número finito de tipos de relações e propõe um elenco inicial que, apesar de não definitivo, pretende ser suficientemente amplo para cobrir a maioria dos casos de inter-relacionamento retórico das proposições do discurso. Ortogonalmente a qualquer tipologia de relações, faz-se uma distinção fundamental entre as relações *hipotáticas*, ou

[1] Projeto CNPq, Proc. Nro. 507030/2004-4.

mononucleares, e *paratáticas*, ou *multinucleares* (Marcu, 1997). As primeiras são necessariamente binárias e estabelecem algum tipo de subordinação entre seus relacionandos, ditos *núcleo* (o principal) e *satélite* (o subordinado). Por sua vez, as relações paratáticas podem ter aridade maior que 2 e estabelecem coordenação entre seus relacionandos, todos ditos núcleos, nesse caso. Assim, em termos arbóreos, as arestas de uma árvore RST são rotuladas de modo a identificar o papel – núcleo (usualmente sob o rótulo *N*) ou satélite (*S*) – de cada subárvore/relacionando. A Figura 1 apresenta um exemplo de árvore RST.

A VT é construída sob a hipótese de uma variante da RST, explorando apenas a oposição hipotaxe-parataxe, ou seja, de forma totalmente ortogonal a qualquer tipologia de relações. Formalmente, a VT ignora os rótulos dos nós (isto é, as próprias relações RST) em favor dos rótulos de arestas. Para qualquer árvore discursiva *t*, a teoria postula uma função $acc_t : edus(t) \rightarrow 2^{edus(t)}$, que permite traçar um mapeamento de cada EDU de *t* para um subconjunto das EDUs de *t*. Como resultado, obtém-se o chamado *domínio de acessibilidade referencial*. A teoria alega, assim, que referências feitas a partir de uma EDU *x* só podem ser resolvidas em $acc_t(x)$, ou seja, no seu próprio domínio de acessibilidade referencial. Desta forma, a c*onjectura C1* passa a ser formalmente definida como segue: para toda árvore *t*, toda EDU $x \in edus(t)$ e toda expressão referencial *e* ocorrendo em *x* uma das seguintes afirmações deve ser verdadeira[2]:

caso I: *e* é *nova no discurso*, ou seja, realiza a primeira menção ao seu referente;

caso II: *e* é anafórica e a primeira menção ao seu referente é realizada em alguma EDU $y \in acc_t(x)$ [3]: [O atacante Nelsinho sofreu uma contusão no jogo do último domingo]$_y$ (...) [*O jogador* deverá ficar fora dos campos por duas semanas]$_x$;

caso III: *e* é anafórica e existem EDUs $y \in acc_t(x)$, $z \in acc_t(y)$ e expressões referenciais $e' \in y$ e $e'' \in z$ tais que *e*, *e'* e *e"* são co-referentes e *e"* é nova no discurso. Nesse caso, *e'* funciona como intermediário para que *e* acesse a menção inaugural *e"*: [Romário entrou na "cruzada" do milésimo gol]$_z$ (...) [*O atacante do Vasco* afirmou que fazer o gol é uma questão de tempo]$_y$ (...) [*O jogador* saiu do Maracanã frustrado]$_x$;

caso IV: a referência em *e* pode ser compreendida na ausência das menções anteriores ao seu referente, como se fosse uma entidade nova no discurso. Os autores denominam esses casos de *referências inferenciais*.

Claramente, os casos problemáticos que são foco de nosso estudo, isto é, que remetem a quebras de cadeias de co-referência, são os casos II e III, os quais denominaremos *casos não-triviais*.

A determinação de acc_t é relativamente simples, pois presta-se a uma implementação direta sobre a topologia de uma árvore RST, como pode ser constatado em (Cristea et al., 1998). A Figura 2 ilustra os acc_t de cada nó da árvore *t* da Figura 1. Nesse grafo de

[2] Que se denota por $e \in x$.
[3] Nesses exemplos, os segmentos sublinhados indicam os referentes e os em itálico, as anáforas.

acessibilidade referencial, os vértices representam EDUs; existe um arco de *x* para *y* se e somente se $y \in acc_t(x)$, ou seja, se a EDU *y* é diretamente accessível a partir de *x*.

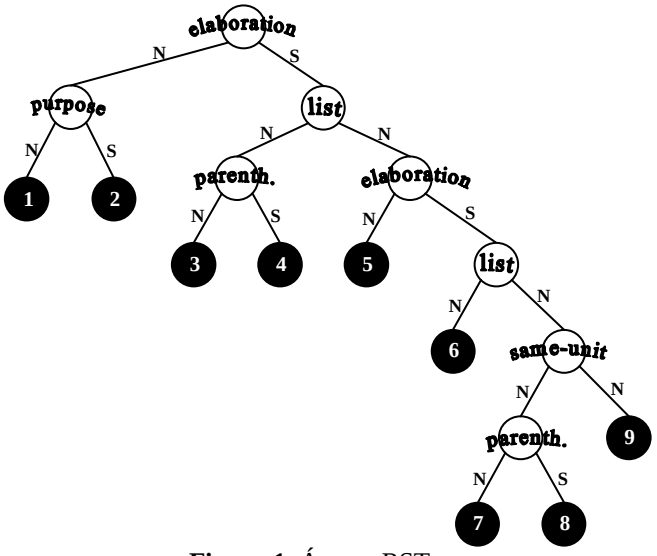

Figura 1. Árvore RST

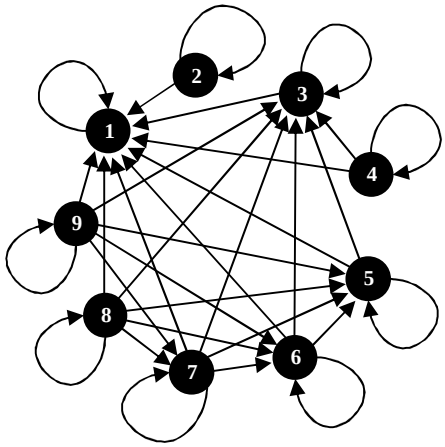

Figura 2. Grafo de acessibilidade referencial para a árvore da **Figura 1**.

3. Metodologia de Validação da Teoria das Veias para o Português

A Figura 3 apresenta um esquema do processo utilizado para chegar aos nossos resultados de validação da C1 para o português.

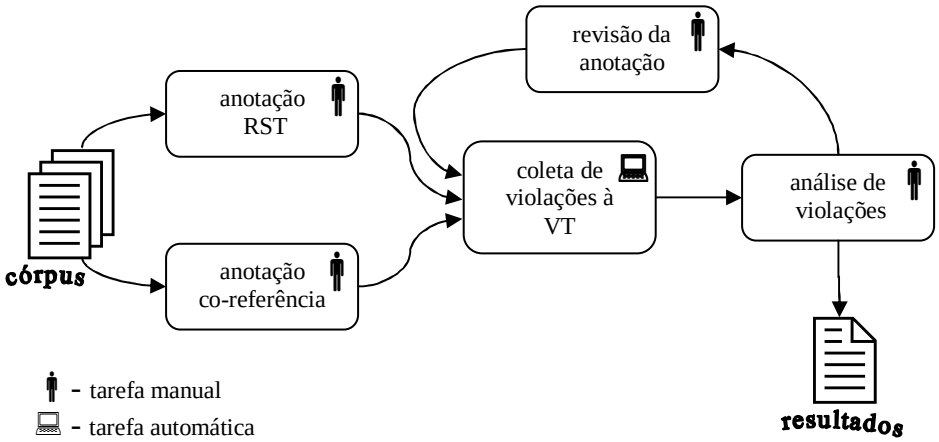

Figura 3. Processo adotado na validação da VT

Primeiramente, constituiu-se o córpus para verificar a validade da C1, o qual foi anotado manualmente quanto à estrutura RST e ao fenômeno da co-referência, por equipes diferentes de lingüistas. As anotações obtidas foram verificadas automaticamente por uma ferramenta que detecta as violações à C1. Neste caso, assume-se que os dados de entrada sejam completos e corretos. Para cada uma das violações acusadas, (i) verificou-se manualmente se a causa não seria algum erro de anotação ou mesmo a inadequação dos

esquemas de anotação e, em caso afirmativo, (ii) procedeu-se às correções possíveis. O corpus assim corrigido foi novamente submetido à detecção de violações e os novos resultados, também depurados. Recorrentemente, essa estratégia foi adotada até que a anotação se estabilizasse. A anotação de co-referência foi realizada com o suporte da ferramenta MMAX (Müller & Strube, 2001); a de RST foi realizada por dois lingüistas com o apoio da ferramenta RSTTool (O'Donnel, 1997). Foram utilizados o elenco de relações de Pardo (2005) e as diretrizes de anotação de Carlson & Marcu (2001).

O córpus utilizado foi um subconjunto do córpus Summit (Collovini et al., 2007), composto por doze textos de divulgação científica do jornal Folha de São Paulo (caderno Ciência), totalizando 3.846 palavras, 430 EDUs e 1.156 expressões referenciais (ERs). Destas, testamos apenas os sintagmas nominais definidos[4], que totalizam 474 itens, referenciados neste artigo por *ERs testáveis*. Elegemos esse recorte por ser a anotação de co-referência uma tarefa dispendiosa, que envolveu esquemas ainda prototípicos, e por julgarmos que, usando essa importante categoria, realizaríamos uma validação preliminar que, além de nos permitir validar e aprimorar nossa metodologia, também produzisse resultados consideráveis.

Fazendo um paralelo com os experimentos de Cristea et al. para o romeno, língua cujo córpus era maior em todas as dimensões consideradas, nosso córpus excede aquele quanto ao número de EDUs (550% maior) e de ERs (941% maior). Entretanto, como lidamos somente com as ERs *testáveis*, o aumento efetivo nessa dimensão é de 327%.

Nosso experimento, no entanto, difere metodologicamente do de Cristea et al. quanto ao cômputo da precisão. Naquele trabalho, jamais se menciona o número de ERs novas no discurso (ERNs) e a precisão é calculada em relação ao número total de ERs. Ora, ERNs são numerosas (no nosso córpus, 68,43% de todas as ERs e 63,50% das testáveis) e correspondem ao caso trivial da C1 (nossos casos I ou IV). Por isso, em nosso experimento, mencionamos o número de ERNs e calculamos a precisão com relação ao total de ERs anafóricas (ERAs), ou seja, aquelas referentes aos casos II ou III. Vale observar que nosso córpus conta com 155 ERAs testáveis, o que ultrapassa em 39,7% o número total de itens testados para o romeno por Cristea et al. Embora tratem-se de experimentos muito díspares, esses números têm por único objetivo aqui ressaltar que nosso experimento preliminar sobre o uso da VT ultrapassa quaisquer outros experimentos antes relatados pelos autores da teoria e não pretendem estabelecer uma comparação efetiva entre ambos.

4. Resultados e Discussão

A Tabela 1 resume os resultados do experimento realizado e inclui, por conveniência, os resultados de Cristea et al aplicados ao inglês, francês e romeno. Os pontos de interrogação indicam os dados que não foram considerados por Cristea et al. Em cada linha, todas as porcentagens são em relação ao número de ERs testáveis, que, no caso de Cristea et al. coincide com o total de ERs. As duas últimas colunas da tabela denotam, respectivamente, o conjunto de testáveis em conformidade com a C1 (**C1-ok)** e o nosso cálculo particular de precisão, dita *Precisão Não-Trivial* (**PNT**). Esta é dada pela seguinte fórmula:

[4] São ditos definidos os sintagmas nominais iniciados por artigo definido, o que inclui algumas ocorrências de nomes próprios.

$$\text{PNT} = 1 - \frac{\left|\overline{\text{C1-ok}}\right|}{\left|\overline{\text{I} \cup \text{IV}}\right|},$$

onde $|\text{X}|$ e $\overline{\text{X}}$ denotam respectivamente o número de elementos do conjunto X e o complemento de X em relação ao universo de testáveis. Intuitivamente, PNT é o complemento de uma taxa de erro mais realista, dada pela razão entre o número de erros – $\left|\overline{\text{C1-ok}}\right|$ – e o total de casos não-triviais – $\left|\overline{\text{I} \cup \text{IV}}\right|$ – cobertos ou não pela C1. Esse cálculo se justifica por sabermos que nenhum dos erros em $\overline{\text{C1-ok}}$ jamais deveria pertencer aos casos I ou IV.

Tabela 1. Cômputo geral dos casos relativos à VT.

				Casos variantes						
Língua	**EDUs**	**ERs**	**Testáveis**	**– I – ERNs**	**– II – Diretas**	**I ∪ II**	**– III – Indiretas**	**– IV – Infer.**	**C1-ok**	**Precisão PNT**
inglês[5]	62	97	97	?	?	75 77,3%	14 14,4%	5 5,2%	94 96,9%	?
francês	48	110	110	?	?	98 89,1%	11 10%	1 0,9%	110 100,0%	?
romeno	66	111	111	?	?	104 93,7%	2 1,8%	5 4,5%	111 100,0%	?
port.	**430**	**1.156**	**474**	**301** **63.5%**	**108** **22.8%**	**409** **86.3%**	**17** **3.6%**	**18** **3.8%**	**446** **94.1%**	**81.94%**

Independentemente de os experimentos nossos diferirem significativamente dos de Cristea et al. e ressaltando que o propósito do experimento reportado não ser uma comparação, mas sim uma validação da VT para o português, fica claro que a precisão daquele conjunto de experimentos é impressionante: apenas 3 ERs não cobertas em 318, todas para o inglês e corrigíveis pela simples conversão de uma relação hipotática em paratática. Entretanto, mais notável ainda é o fato de que a proximidade lingüística do português com o francês e o romeno não fique aí refletida. Isso parece sugerir que diferenças lingüísticas não devem ser responsáveis pelo contraste observado. Antes, cremos que tenham sido determinantes diferenças relativas a (i) gêneros textuais no córpus e (ii) esquema de anotação RST.

Quanto a gêneros textuais, podemos afirmar que, pelo menos para o romeno e o francês, foram usados fragmentos de narrativa; o gênero dos três textos em inglês não foi mencionado. Nossos textos de divulgação científica possuem estrutura retórica sabidamente distinta dos narrativos. Entretanto, cabe lembrar que há carência de estudos mais aprofundados sobre o impacto do gênero sobre a estrutura RST, especialmente sobre sua topologia e distribuição da nuclearidade, dois fatores importantíssimos para a VT. Quanto aos esquemas de anotação RST, há grande variação tanto da tipologia das relações quanto das diretrizes de segmentação em EDUs, interferindo na anotação diretamente. Infelizmente, Cristea et al. não especificam qual esquema usam. Nesse sentido, também nosso experimento prova ser mais completo do que o deles. Finalmente, a inovação de usar a PNT se prova válida ao desfazer qualquer otimismo com a precisão geral de 94,1%

[5] Dados reproduzidos de Cristea et al. (1998) para fins de comparação.

observada (vide coluna **C1-ok**). Um poder preditivo real de 81,94% certamente sugere melhorias e revela uma C1 bem menos absoluta que a reportada originalmente por Cristea et al.

5. Depuração e Lições Aprendidas

Toda tarefa de anotação de córpus está sujeita a erros e todo esquema de anotação faz um recorte fenomenológico que pode se provar insuficiente em certas situações, para não mencionar a própria possibilidade de conter erros conceituais. Durante a depuração dos nossos resultados, encontramos todas essas situações em ambas as modalidades de anotação utilizadas.

Quanto à anotação RST, contamos seis erros comuns de anotação, ora de segmentação, ora de inversão de nuclearidade (trocar núcleo por satélite e vice-versa). Houve mais 7 casos de falha devido à relação *attribution*, muito comum no gênero jornalístico. Segundo Carlson & Marcu, na relação *attribution* (hipotática), o núcleo apresenta a expressão, fala ou pensamento de alguém, ao passo que o satélite indica o respectivo emissor. Em nossas anotações, chegamos a considerar falha conceitual de nosso esquema, como ilustra o exemplo E1:

E1: ["Em oito anos, detectamos mais de 300 eventos, graças ao nosso sistema de calibragem dos dados de satélite"]$_N$ [, *conta Douglas Revelle, do Laboratório Nacional de Los Alamos, um dos autores do estudo, que está publicado na edição de hoje da revista britânica "Nature" (www.nature.com).*]$_S$

Observamos que, no texto jornalístico, é extremamente comum a introdução de novos referentes no satélite de relações *attribution*. Além disso, é usual que esses referentes sejam retomados posteriormente no discurso. Entretanto, é um corolário da VT que jamais um satélite *S* numa subárvore de raiz *R* pertencerá ao *acc* de qualquer nó cujo caminho para *S* passe por nós acima de *R*. Assim, é muito comum que as referências posteriores não satisfaçam a C1. Esse problema pode ser evitado se considerarmos todas as relações *attribution* como paratáticas (multinucleares). Isso permitirá o acesso a *S* pelo menos pela subárvore de todo ancestral *R'* de *R* tal que só haja arestas N separando *R* de *R'*.

Quanto à anotação de co-referência, houve três casos de erro trivial de anotação, de marcação de ER nova no discurso com anafórica. Por outro lado, no que concerne a deficiências conceituais, o quadro é um pouco mais complexo do que para RST. Em específico, nosso esquema se concentra numa co-referência estrita, sem explicitar a possibilidade de resolução de certas ERs na ausência de suas ERs co-referentes precedentes (Casos I ou IV). Referências inferenciais (Caso IV) constituem o caso mais freqüente dessa situação. Veja os exemplos seguintes:

E2: "... [o País]$_{i,nova}$ ... [o Brasil]$_{i,velha}$ ..."

E3: "... [o fígado]$_{j,nova}$ ... [células de [o fígado]$_{j,velha}$]$_{k,nova}$ ..."

Nos exemplos E2 e E3, temos as ERs "o Brasil" e "o fígado" (2ª ocorrência), claramente interpretáveis na ausência de seus antecedentes, marcadas como ERs anafóricas quaisquer pela simples razão de nosso esquema de anotação não distinguir esses casos.

Em termos de revisão do esquema de anotação (e não de custo de anotação ou mesmo reprodutibilidade desta por computador), a solução para os casos de ERs inferenciais é trivial, podendo ser efetuada pela mera adição de um traço de anotação. Entretanto, existem diversos outros exemplos menos claros, cujo tratamento deixamos para trabalhos futuros, como os seguintes:

E4: "[Um ser que invade corpos e domina a mente alheia, forçando suas vítimas a fazer o que ele ordena,]$_{i,nova}$ não é mero personagem de ficção. Para uma aranha da Costa Rica, essa criatura existe ... Apesar do nome *Hymenoepmecis sp.*, [o tal invasor de corpos]$_{i,velha}$ é só [uma vespa]$_{i,velha}$ ..."

Em E4, a ER "uma vespa" é anotada como anafórica por sua co-referência com ERs anteriores, as quais se encontram em satélites que não estão acessíveis a ERs posteriores verdadeiramente dependentes de "uma vespa". Estas ERs são consideradas, então, como violações à C1, apesar de acessarem a EDU onde se encontra "uma vespa".

E5: "... poderiam originar [as células hepáticas]$_{j,velha}$, além de [as sangüíneas]$_{k,velha}$"

Temos em E5, um caso curioso, em que a ER "as sangüíneas" não é co-referente com "as células hepáticas", mas depende desta para ser interpretada. Esse tipo de dependência não é capturado por nosso presente esquema de anotação.

5. Considerações finais

Reportamos neste artigo o primeiro esforço de validação da C1 para o português. Trata-se de um experimento preliminar, na medida em que cobrimos apenas uma determinada categoria de sintagmas nominais anafóricos, a saber: os introduzidos por artigo definido. Apesar dessa limitação, obtivemos resultados interessantes tanto para (i) aperfeiçoar a metodologia de validação para um experimento mais completo quanto para (ii) observar discrepâncias significativas em relação aos experimentos similares já realizados para outras línguas. Quanto a (i), coletamos evidências das limitações dos esquemas de anotação utilizados e introduzimos um índice de validação – a Precisão Não-Trivial – que parece estimar de forma mais realista o poder preditivo da C1. Quanto a (ii), observamos resultados significativamente piores que os de Cristea et al, provavelmente devido a diferenças de córpus e esquemas de anotação.

Como trabalhos futuros, esperamos (i) complementar e revisar a anotação atual do nosso córpus para possibilitar um experimento completo, (ii) estudar mais detidamente cada caso de violação da C1, em busca de alguma generalização ou até uma revisão da teoria, e (iii) investigar as causas de discrepância entre os experimentos.

Agradecimentos

Este trabalho conta com o apoio do CNPq e CAPES.

Referências bibliográficas

Carlson, L.; Marcu, D. (2001). Discourse Tagging Reference Manual. ISI Technical Report ISI-TR-545.

Collovini, S.; Carbonel, T. I.; Fuchs, J. T.; Coelho, J. C.; Rino, L.; Vieira, R. (2007). Summ-it: um córpus anotados com informações discursivas visando à sumarização automática. Anais do V Workshop em Tecnologia da Informação e da Linguagem Humana – TIL'2007. Rio do Janeiro-RJ. Julho.

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

Grosz, B.; Joshi, A.; Weisten, S. (1995). Centering: A Framework for Modeling the Local Coherence of Discourse. *Computational Linguistics*, V. 21, N. 2, pp. 203-225.

Mann, W.C.; Thompson, S.A. (1987). Rhetorical Structure Theory: A Theory of Text Organization. Technical Report ISI/RS-87-190.

Marcu, D. (1997). The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts. PhD Thesis, Department of Computer Science, University of Toronto.

Marcu, D. (1999). A formal and computational synthesis of Grosz and Sidner's and Mann and Thompson's theories. In the Proceedings of the Workshop on Levels of Representation in Discourse, pp. 101-108. Edinburgh, Scotland.

Müller, C.; Strube, M. (2001).Annotating anaphoric and bridging relations with mmax. In Proceedings of the 2nd SIGdialWorkshop on Discourse and Dialogue, Aalborg, Denmark, pp. 90-95.

O'Donnel, M. (1997). RSTTool: An RST Analysis Tool. In Proc. of the 6th European Workshop on Natural Language Generation, Gerhard-Mercator University, Duisburg, Alemanha.

Pardo, T.A.S. (2005). Métodos para Análise Discursiva Automática. Tese de Doutorado. Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. São Carlos-SP, Junho, 211p.

Seno, E.R.M. (2005). Especificação de Heurísticas de Sumarização de Estruturas RST com Base na Preservação dos Elos Co-Referenciais. Dissertação de Mestrado. Departamento de Computação, UFSCar.

Seno, E.R.M., Rino, L.H.M. (2005). Co-referential chaining for coherent summaries through rhetorical and linguistic modeling. In: Proceedings of Workshop on Crossing Barriers in Text Summarization Research – RANLP´05. p.70-75.