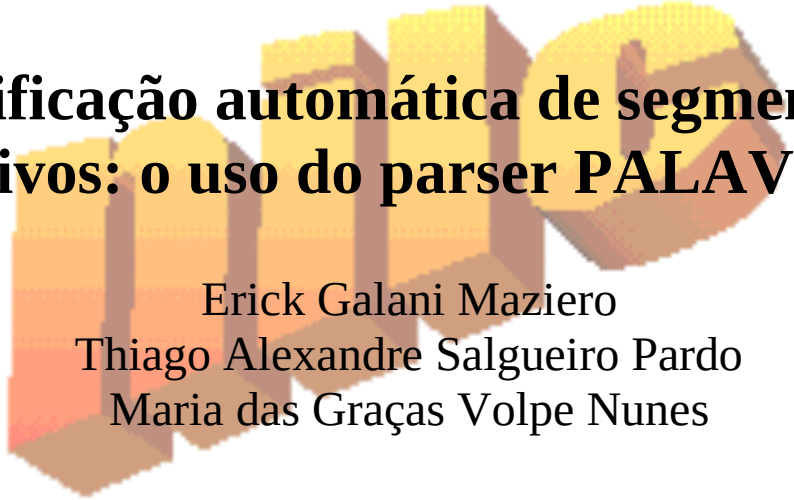
Universidade de São Paulo - USP
Universidade Federal de São Carlos - UFSCar
Universidade Estadual Paulista - UNESP



Identificação automática de segmentos discursivos: o uso do parser PALAVRAS

Erick Galani Maziero
Thiago Alexandre Salgueiro Pardo
Maria das Graças Volpe Nunes

NILC-TR-07-08

Agosto, 2007

Série de Relatórios do Núcleo Interinstitucional de Lingüística Computacional
NILC - ICMC-USP, Caixa Postal 668, 13560-970 São Carlos, SP, Brasil

Resumo

O processo de segmentação textual é uma tarefa prévia para a maior parte das aplicações de Processamento de Língua Natural (PLN), sendo que tarefas diferentes exigem segmentos com granularidades diferentes. Este trabalho visa à produção de segmentos que encerrem em si uma idéia ou conceito básico do texto, os quais são ideais a uma análise retórica/discursiva do texto. Mais especificamente, aborda-se a RST (*Rhetorical Structure Theory*), uma das teorias discursivas mais utilizadas atualmente. O método de segmentação aqui exposto será incorporado ao sistema DiZer (*DIscourse analyZER for BRazilian Portuguese*), substituindo sua etapa de segmentação textual, objetivando melhor desempenho deste analisador retórico automático pioneiro para o português do Brasil. O método apresentado baseia-se em informações morfossintáticas produzidas pelo parser PALAVRAS, um dos melhores analisadores para o português do Brasil.

ÍNDICE

1. Introdução

O Processamento de Línguas Naturais (PLN) engloba uma grande quantidade de subáreas, dentre as quais podemos citar: a Sumarização Automática (SA), que visa à obtenção de resumos de textos automaticamente; a Tradução Automática (TA), que, dado um texto numa língua de origem, busca sua melhor tradução em uma língua alvo; e a Análise Discursiva Automática (ADA), que produz estruturas que refletem os objetivos e a organização do texto.

A maioria das tarefas em PLN, ao processar o texto inicialmente, não trabalha com todo o texto simultaneamente, mas processa partes dele. Essas partes menores do texto, chamadas de segmentos do texto, não têm as mesmas características para todas as aplicações. Por exemplo, para realizar uma tarefa de TA, pode-se optar por trabalhar com segmentos equivalentes a palavras do texto (embora isto não produza bons resultados). Em ADA, não faz muito sentido estabelecer relações entre as palavras de um texto; opta-se, então, por segmentos com características próprias para esta tarefa.

Em ADA, pressupõe-se que um texto contém uma estrutura bem elaborada, onde cada uma de suas partes não é desconexa ou desordenada, mas segue padrões, de acordo com os objetivos do compositor do texto. Assim, trabalha-se para produzir uma estrutura, muitas vezes chamada retórica, no formato de uma árvore, em que cada folha desta árvore é um trecho do texto (segmentos), esses trechos não se sobrepõem, e os nós são relações que se estabelecem entre os trechos de texto ligados a este nó. Além desta caracterização tem-se a definição de trechos mais importantes em uma relação; um trecho é núcleo (mais importante) e tem como um complemento um trecho satélite (menos importante). Isso é exatamente o que prescreve a RST (*Rhetorical Structure Theory*) (Mann e Thompson, 1987), umas das teorias discursivas mais difundidas atualmente e foco deste trabalho.

Há vários métodos para se obter os segmentos de um texto: os baseados em expressões regulares, como o de Walker et al. (2001), que usa padrões de busca, isto é, regras fixas para localizar os extremos das sentenças do texto; há os que utilizam análise numérica e estatística, como o sistema *TextTiling* de Palmer e Hearst (1994, 1997), que busca uma segmentação topical do texto, ou seja, dividir o texto em tópicos, ou porções que tratam de um mesmo assunto; e os métodos que se utilizam de Aprendizado de Máquina, uma subárea da Inteligência Artificial, como o sistema *Satz* de Palmer e Hearst (1994, 1997) para delimitação de sentenças do texto. Mais detalhes sobre métodos de segmentação podem ser encontrados em Pardo e Nunes (2002).

O método aqui proposto caracteriza-se por utilizar regras baseadas em informação morfossintática provenientes do parser PALAVRAS (Bick, 2000), um dos melhores analisadores sintáticos automáticos para o português do Brasil. Visa-se à obtenção de segmentos para posterior incorporação do analisador discursivo automático DiZer (*DIscourse analyZER for Brazilian Portuguese*) (Pardo, 2005), o qual é baseado na RST. O processo atual de segmentação do DiZer baseia-se em regras simples, sujeitas a muitos erros. Por outro lado, sabe-se que a sintaxe tem um papel importante no processo de segmentação, como pode ser evidenciado no trabalho de Carlson e Marcu (2001). Este relatório descreve o processo de segmentação desenvolvido, que, futuramente, deverá ser acoplado ao DiZer.

A seguir, na Seção 2, apresentam-se as noções básicas de segmentação para fins de análise discursiva. Na Seção 3, o processo completo de segmentação é delineado, enquanto as regras de segmentação desenvolvidas são apresentadas na Seção 4. Algumas considerações finais são apresentadas na Seção 5.

2. Identificação de segmentos discursivos

Como os segmentos de um texto para análise discursiva têm suas peculiaridades, foi feito um estudo do documento apresentado por Carlson e Marcu (2001), que lista algumas regras para a segmentação do texto objetivando a análise RST. Embora esse documento tenha sido elaborado para a língua inglesa, tomou-se o cuidado de considerar apenas as regras que se aplicam à língua portuguesa. Esse trabalho foi a base para o desenvolvimento do segmentador proposto.

Define-se a Unidade Básica do Discurso (UBD) (tradução de *elementary discourse unit*, termo utilizado por Carlson e Marcu) como sendo uma parte (segmento) do texto que encerra em si uma idéia ou conceito básico. Considere a porção de texto:

> "*No centro do método está um gene humano descoberto recentemente, o UCP-3, cujos mecanismos de ação ainda não são totalmente conhecidos.*"

Uma palavra não pode ser considerada uma UBD, pois, isolada, não traz uma idéia completa, ou proposição, no texto. Por exemplo, ao considerar a palavra *centro*, não temos percepção de um significado no cenário do texto. No entanto, a sentença enunciada acima contém mais de uma idéia básica; temos que sentenças, ou períodos, podem conter mais de uma UBD; no caso acima, temos 2 UBDs, sendo que a primeira fala do gene descoberto e a segunda afirma que os mecanismos de ação deste gene ainda não são totalmente conhecidos. Podemos, então, definir as seguintes UBDs:

> 1: *No centro do método está um gene humano descoberto recentemente, o UCP-3,*
> 2: *cujos mecanismos de ação ainda não são totalmente conhecidos.*

Em geral, uma proposição, ou unidade mínima de significado, é expressa por uma oração. Se fossem utilizados métodos para delimitar sentenças do texto, como Walker et al. (2001), o processo de análise retórica não teria o máximo de aproveitamento, pois não conseguiria expor a relação que há entre idéias expressas dentro de uma mesma sentença. Percebemos claramente que uma segmentação topical pioraria ainda mais o desempenho da análise retórica (pelo menos como predito na RST) por conter muitos conceitos em cada tópico, dependendo de seu tamanho.

O manual desenvolvido por Carlson e Marcu (2001) indica como obter segmentos que expressem conceitos básicos do texto para sua inter-relação. Marcu (2000), identifica as relações entre os segmentos pela identificação de marcadores discursivos, ou seja, expressões lingüísticas que indicam a estrutura discursiva subjacente ao texto, como o "mas" no exemplo abaixo:

> "*Por enquanto é só sugestão: o tratamento foi testado em camundongos.* ***Mas*** *os resultados levaram os cientistas a chamar o próprio estudo de abordagem promissora contra a obesidade.*"

O marcador discursivo "***Mas***" no texto relaciona o texto que o antecede com o procedente estabelecendo uma relação de contradição entre seus significados.

O método de segmentação aqui exposto também se baseia na localização e papel desempenhado por esses marcadores discursivos no texto para criação das regras de segmentação.

A seguir, o método proposto é delineado.

3. Segmentação textual

Antes de se segmentar o texto, este é submetido ao parser PALAVRAS (Bick, 2000), que o analisa morfossintaticamente, colocando, para cada palavra, marcações sobre sua classe morfológica e seu papel sintático (essas marcações são listadas mais abaixo).

Analisando um corpus anotado retoricamente por um especialista da área, foram criadas regras para a segmentação automática, verificando-se onde ocorriam as delimitações de segmentos e buscando suas equivalentes marcas morfossintáticas. Foram levadas em consideração as regras do manual de segmentação de Carlson e Marcu (2001). As regras desenvolvidas são ilustradas na próxima seção.

Para a segmentação, preliminarmente, é feito um processamento do texto original pelo parser Palavras obtendo-se as marcações morfossintáticas. O texto resultante do processamento acima é passado por um parser secundário que converte as marcações de interesse num formato que facilitará a segmentação textual. Este parser secundário gera marcas morfológicas e sintáticas também para cada palavra e realiza a segmentação sentencial.

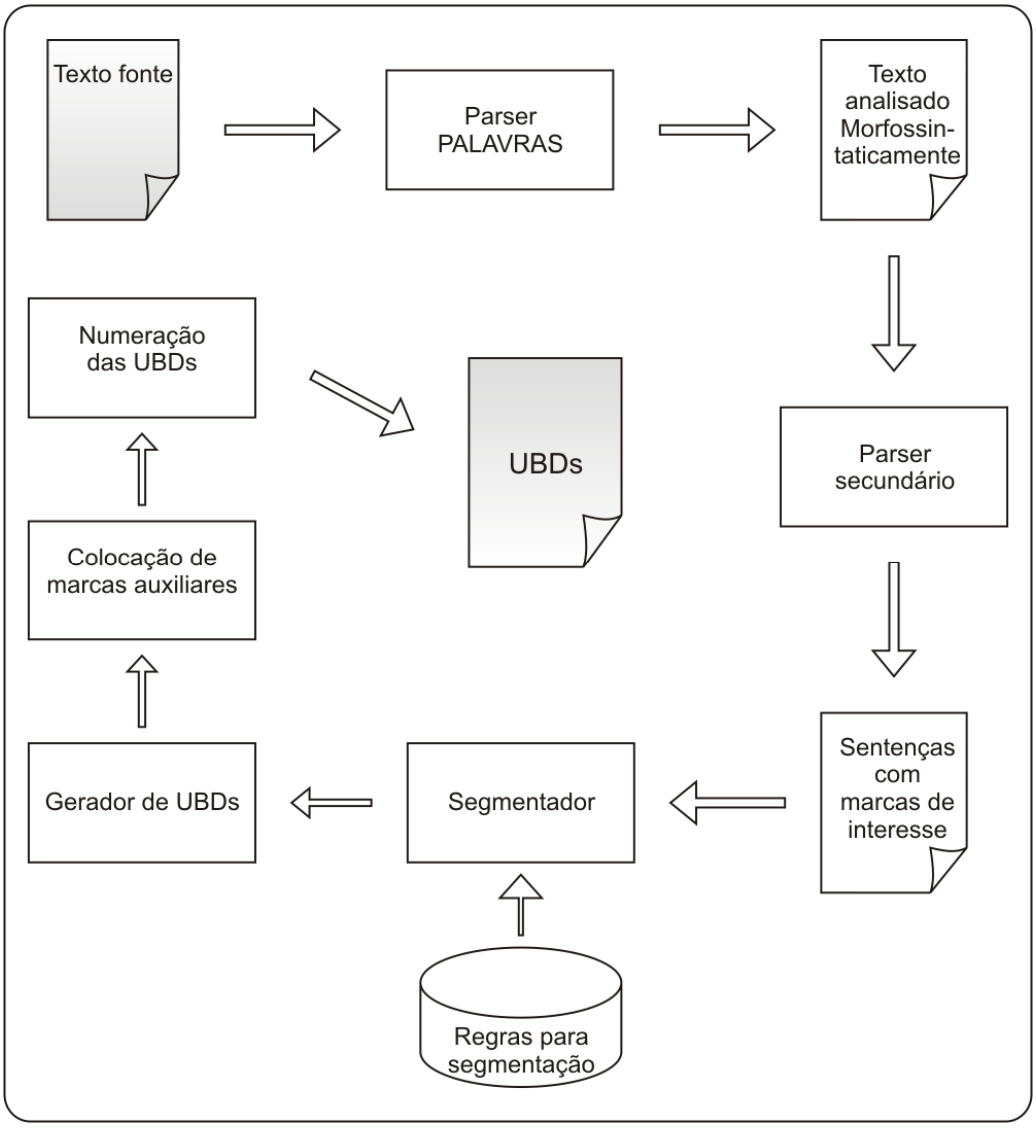


Figura 1 – Passos da segmentação

Então, para cada sentença, ou período, são aplicadas as regras desenvolvidas gerando segmentos que são listados em um arquivo de saída, um por linha.

Temos que as UBDs devem conter no mínimo um verbo (para que se constitua uma oração), com exceção das UBDs acopladas (traduzido da palavra em inglês *embedded*) (trechos de texto que aparecem dentro de uma UBD, mas não constituem uma idéia básica do discurso, realizando o papel de assessorar uma UBD). Assim, é verificado para cada segmento gerado se este contém algum verbo; caso não contenha, o segmento posterior é juntado a este, produzindo um novo segmento.

Algumas marcações são deixadas no texto resultante do processo de segmentação, de forma a facilitar o processo do DiZer, tais como indicadores de verbos de atribuição e segmentos entre parênteses (UBDs acopladas), necessários para a identificação de algumas relações retóricas no sistema citado. Os segmentos são então numerados sequencialmente, um segmento por linha.

A Figura 1 ilustra esse processo. Detalhes de execução, tais como comandos para execução do parser PALAVRAS, são mostrados no Apêndice B.

O segmentador desenvolvido encontra-se disponível *on-line* para uso pela comunidade de pesquisa na página do NILC[1], grupo ao qual este trabalho pertence. As Figuras 2 e 3 ilustram a interface do segmentador antes e após a segmentação de um texto de exemplo, sendo que os detalhes observados nos resultados são explicados na próxima seção.

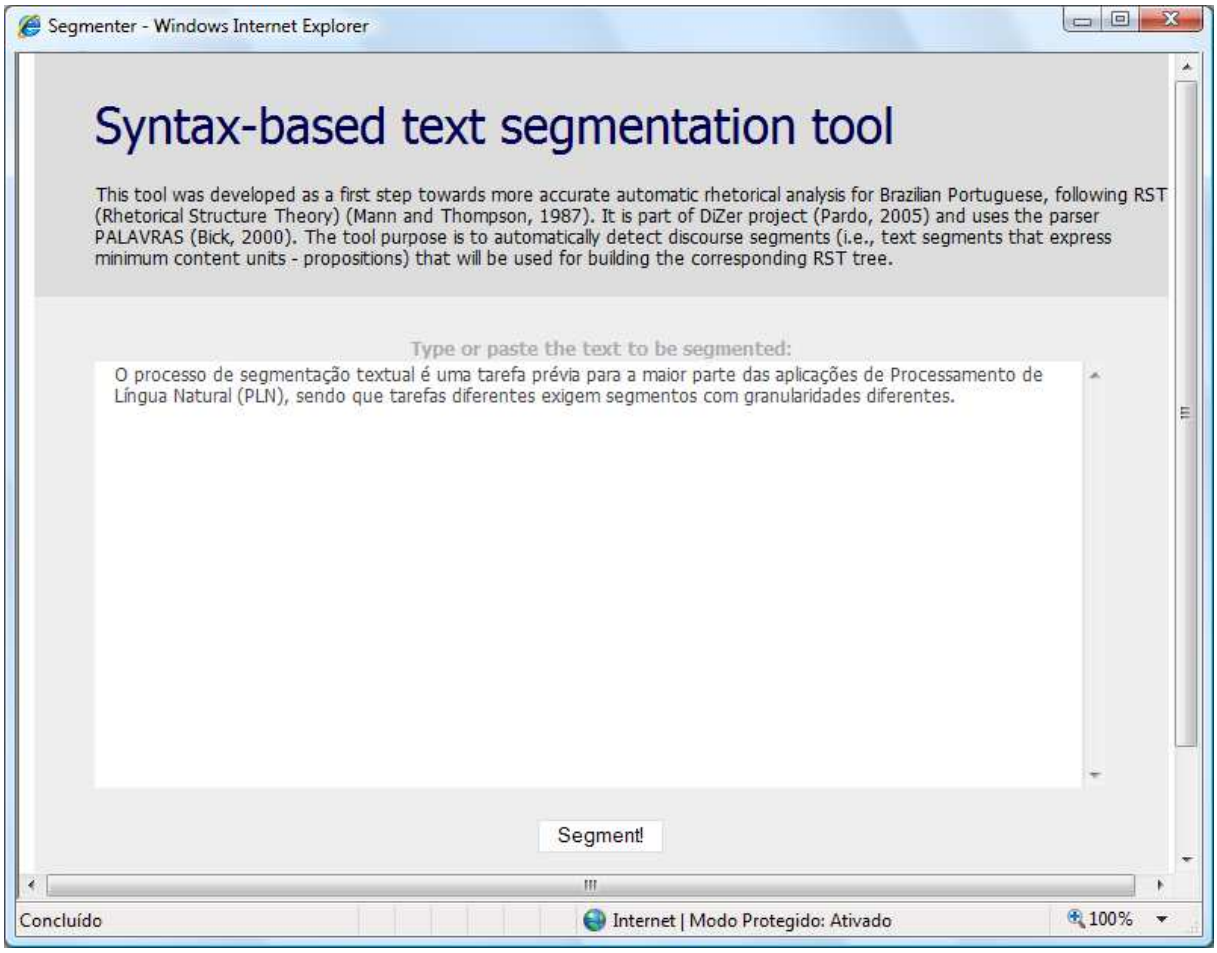


Figura 2 – Interface *on-line* de segmentação (antes do processo começar)

[1] www.nilc.icmc.usp.br

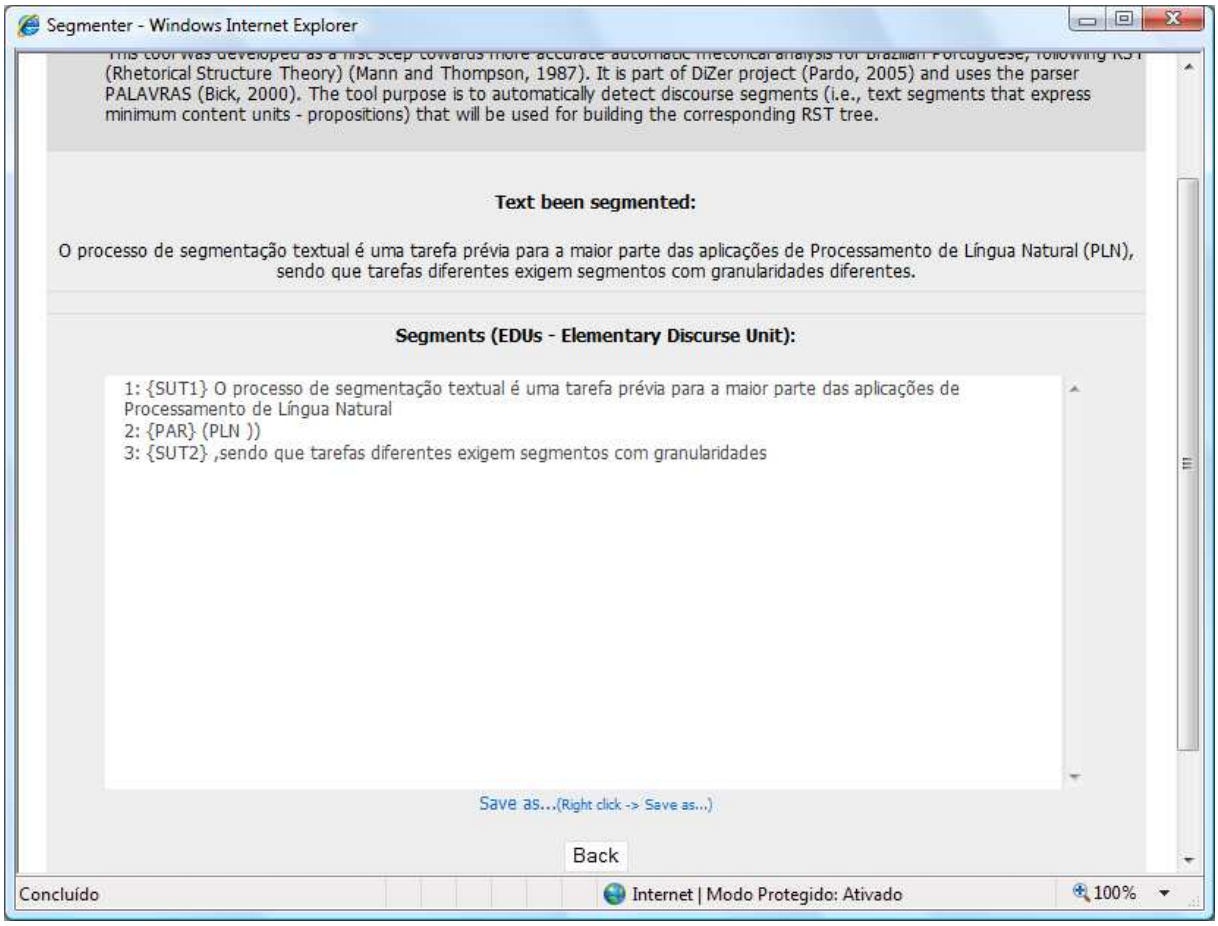


Figura 3 – Interface *on-line* de segmentação (ao fim do processo)

A seguir, as regras de segmentação desenvolvidas são apresentadas.

4. As regras de segmentação

Listam-se, abaixo, as regras produzidas com exemplos de aplicação e de exceção. Considere cada segmento delimitado por colchetes. Note que os segmentos das regras superiores podem conter outros segmentos que serão obtidos com regras mais abaixo. Mostra-se, no Apêndice A, um texto completo sendo segmentado.

Regra 1 – segmenta-se um trecho de texto ao encontrar a marca </s>, dada pelo parser PALAVRAS, sendo que essa marca é posta para indicar o fim de uma sentença;

Exemplo:
[O índice está abaixo de média para o horário, que é de 119 km -em comparação a quinta da semana passada.] [O motorista encontra 9 km de lentidão na marginal Pinheiros , na pista local , sentido, da rua Doutor Rubens Gomes Bueno até a ponte Eusébio Matoso.]

Regra 2 - os trechos entre parênteses, chaves, colchetes ou traço;

Exemplo:

[No corredor norte-sul] [(avenidas 23 de Maio, Rubem Berta e Moreira Guimarães),] [sentido aeroporto , há lentidão entre a praça da Bandeira até o viaduto Indianópolis.]

Regra 3 – após os verbos atributivos;

Exemplo:
[Esse é o pior panorama climático previsto pelo instituto,] [disse Carlos Nobre, que participou do debate "Cenários da Amazônia", na 52ª Reunião Anual da SBPC.]

Regra 4 – antes de verbo no gerúndio ou particípio, quando este vem precedido de vírgula;

Exemplos:
[A larva passa de 7 a 14 dias ali dentro,] [fartando-se do sangue do aracnídeo,] [até estar madura o suficiente.]
[A análise do DNA dessas células mostrou que elas continham o cromossomo Y,] [encontrado apenas em células masculinas.]

Regra 5 – antes de conjunções coordenativas com a marca de desambiguação <co-fmc> ou <co-fin> que não sejam procedidas por objetos acusativos ou verbos;

Exemplo:
[Tudo o que sobrou dele foram as patas e partes da cauda e da bacia,] [mas os pesquisadores conseguiram estimar que o bicho fosse um filhote de 1,5 metro de altura.]

Exceção:
["É uma descoberta] [e tanto",] [disse o psicólogo César Ades, da USP, especialista em comportamento de aranhas.]

Regra 6 – antes de conjunções coordenativas precedidas por vírgulas ou ponto e vírgula;

Exemplo:
["Mães americanas tendem a usar mais substantivos,] [e a usar mais objetos ao brincar com seus filhos pequenos.]

Regra 7 – antes de conjunções subordinativas precedidas por vírgulas ou ponto e vírgula;

Exemplo:
[Poynar pretende usar esse material,] [que segundo ele tende a ter aparência e consistência de chocolate...]

Exceção:
[Para ele,] [se um paradigma experimental ou analítico não for usado de alguma forma em um projeto, este provavelmente não será um projeto de pesquisa.]

Regra 8a – antes de orações relativas, marcadas por <rel> e que sejam advérbios ou especificadores (SPEC), não precedidas por verbos ou objetos acusativos;

Exemplo:

[...disse Carlos Nobre,] [que participou do debate "Cenários da Amazônia", na 52ª Reunião Anual da SBPC.]

Exceção:
[Os pesquisadores argumentaram que as mulheres] [que não receberam AZT não o teriam recebido,...]

Regra 8b – item 8 precedido por determinante;

Exemplo:
[A intenção do governo é usar parte da soja transgênica já plantada no país,] [e que está com seu consumo proibido, na produção do combustível.]

Exceção:
[Adalberto Veríssimo, da ONG Imazon, apresentou estudo segundo] [o qual as cidades em regiões amazônicas ocupadas de forma predatória duram por volta de 23 anos.]

Regra 9 – Advérbios de relação ou preposições marcadas com ADL, precedidos por vírgula ou ponto e vírgula ;

Exemplo:
[Depois disso, entram em colapso total,] [por falta de uma política de desenvolvimento sustentável.]

Exceção:
[Por severa escassez de água potável, entende-se,] [segundo a ONU, ...]

Regra 10 – Palavras adverbiais com marca de desambiguação <kc> precedidas por conjunção subordinativa;

Exemplo:
[Poynar pretende usar esse material,] [que segundo ele tende a ter aparência e consistência de chocolate, ...]

Regra 11a – expressões apositivas;

Exemplo:
[...disse a secretária de Coordenação da Amazônia do Ministério do Meio Ambiente,] [Mary Allegretti.]

Regra 11b - expressões apositivas cuja palavra com marca de apositivo é precedida por determinando ou numeral;

Exemplo:
[No centro do método está um gene humano descoberto recentemente,] [o UCP-3,] [cujos mecanismos de ação ainda não são totalmente conhecidos.]

Regra 12 - Não se segmenta dentro dos trechos descritos no item 2, acima.

As marcações empregadas pelo PALAVRAS (algumas já citadas anteriormente) são listadas abaixo. Algumas foram transformadas pelo parser secundário em outras marcas com fins de facilidade de entendimento e processamento. Tais transformações também são indicadas abaixo.

- marca geral:

  </s> - indica fim de sentença, transformada em fim de linha no arquivo.

- marca morfossintática:

  <-sam> - no caso de palavras compostas por preposição e artigo (que foram separadas, como no caso de "da", que o PALAVRAS tratou como "de a"), marca uma preposição, indicada por PRP.
  PERS – indicando pronomes pessoais, transformada na marca PRO.
  PRP – Indicando preposições
  V GER – verbo no gerúndio, transformada em Verbger.
  V PCP – verbo no particípio, transformada em Verbpcp.
  V @FS-STA – indicando oração principal finita, transformada em Verbfsta.
  V <fmc> – verbo de oração principal finita, transformada em Verbfmc.
  V – verbo, transformada em Verb. (Essa marca será transformada somente se as possibilidades de verbos acima não forem identificadas antes).
  ADV <rel> - advérbio de relação, transformada em ADVrel.
  ADV - advérbio.
  ADJ – adjetivo.
  N – substantivo, transformada em NOUM.
  PROP – substantivo próprio, transformada em PN.
  KC – Conjunção coordenativa.
  KS – Conjunção subordinativa.
  SPEC <rel> – pronomes relativos independentes, transformada em SPECrel.
  SPEC – palavras definidas como pronomes não flexionados.
  DET – determinantes, definidos como pronomes flexionados.
  NUM – numerais.

- marcas sintáticas:

  @SUB – indica subordinação, transformada em CS.
  @CO <co-fmc> - indica coordenação, onde <co-fmc> indica o que é coordenado, neste caso a oração principal finita, essa marca foi transformada em COfmc.
  @CO <co-fin> - indica coordenação, onde o verbo finito é coordenado, marca transformada em COfin.
  @CO - indica coordenação.
  APP – indica expressão apositiva, sempre aparece após vírgula.
  @SUBJ – indica sujeito da oração, transformada em SBJ.
  ACC – indica objeto direto (acusativo).
  ADVL <ks> – advérbio conjuncional, transformada em ADLkc.
  ADVL <advl-rel> - advérbio de relação, transformada em ADLrel.
  ADVL – advérbio, transformada em ADL.

Para tornar a análise efetuada pelo DiZer mais eficiente, também são feitas as seguintes marcações no arquivo final do processo de segmentação:

- Marca com {PAR} o segmento de relação da RST *parenthetical* (segmentos acoplados). Caso esse segmento esteja no meio de um segmento, que então só foi segmentado devido ao segmento *parenthetical*, são adicionadas as marcas

{SUT1} e {SUT2} para os segmentos anteriores e posteriores, respectivamente, ao marcado por {PAR}.

Exemplo:
[{SUT1} O Instituto Nacional de Pesquisas Espaciais] [{PAR} (Inpe)] [{SUT2} prediz um aumento de temperatura de até 5C nas áreas mais secas da Amazônia , em 50 anos ...]

- A exemplo do que ocorre no parágrafo anterior, foram marcados segmentos apositivos com a marca {APP}, para serem tratados como segmentos acoplados, e que portanto podem quebrar segmentos. Quando a quebra de um segmento ocorre com a expressão apositiva no meio, os dois trechos de segmentos são marcados com {SUT1} e {SUT2}, como no caso da relação *parenthetical* acima.

  Exemplo:
  [{SUT1] O presidente da Comissão Nacional de Ética em Pesquisa,] [{APP} William Saad Hossne,] [{SUT2} disse ontem, na 52ª Reunião Anual da Sociedade Brasileira para o Progresso da Ciência,]

- Para identificação de complementos de verbos de atribuições, que são considerados segmentos, foram usadas as marcas {ATA} ou {ATB}. A marca {ATB} indica que o complemento do verbo atributivo encontra-se no segmento após o segmento do verbo de atribuição. Já {ATA} indica que o complemento encontra-se antes.

  Exemplo:
  [Ele afirmou] [{ATA} que o aumento de temperatura,]...
  [Os dois fenômenos climáticos combinados levariam à desertificação de algumas áreas,] [{ATB} disse ele.]

Algumas considerações finais são feitas a seguir.

5. Considerações finais

O processo apresentado mostrou-se eficiente na determinação de UBDs (Unidades Básicas do Discurso) dada a utilização de conhecimento morfossintático, permitindo ir além da segmentação oracional ou sentencial.

Os segmentos gerados podem ser usados em análises retóricas tanto manuais quanto automáticas; o uso em análises manuais permite maior conformidade entre os analisadores de um mesmo texto, pois os segmentos utilizados para gerar a árvore retórica serão os mesmos.

Como trabalhos futuros, pretende-se acoplar o processo de segmentação desenvolvido ao DiZer e realizar uma avaliação completa do procedimento, isolado e em conjunto com a análise discursiva automática.

Agradecimentos

Este trabalho contou com o apoio das agências de fomento à pesquisa FAPESP, CAPES e CNPq.

Referências

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

Eckhard B. (2000). *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Aarhus University Press.

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

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

Palmer, D. D. and Hearst, M.A. (1994). Adaptive Sentence Boundary Disambiguation. In the *Proceedings of the Conference on Applied Natural Language Processing*. Stuttgart, Germany.

Palmer, D. D. and Hearst, M.A. (1997). Adaptive Multilingual Sentence Boundary Disambiguation. *Computational Linguistics*, Vol. 23, No. 2, pp. 241-267.

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.

Reynar, J.C. (1999). Statistical Models for Topic Segmentation. In the *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pp. 357-364.

Walker, D.J.; Clements, D.E.; Darwin, M. and Amtrup, J.W. (2001). Sentence Boundary Detection: A Comparison of Paradigms for Improving MT Quality. In the *Proceedings of the 8th Machine Translation Summit*.

Apêndice A:
exemplo integral das etapas do processo de segmentação

Texto original:

Pesquisadores do Museu Nacional do Rio de Janeiro anunciaram ontem a descoberta de uma nova espécie de dinossauro no Brasil. O animal era um carnívoro que habitou o nordeste brasileiro há 110 milhões de anos, no período Cretáceo (o último da era dos grandes répteis).

Batizado de Santanaraptor placidus, o fóssil é o único a ser encontrado no país com restos de tecido mole, como fibras musculares, vasos sanguíneos e pele.

Para os paleontólogos, achar esse tipo de evidência equivale a acertar na loteria.

"É como se o dinossauro tivesse sido enterrado ontem", disse Alexander Kellner, geólogo do Setor de Paleovertebrados do Museu Nacional e coordenador da expedição que encontrou o fóssil na região da Chapada do Araripe, Ceará (veja mapa).

Com os tecidos preservados, os cientistas esperam poder saber mais sobre o modo de vida e a evolução dos répteis.

Outra importante descoberta é que, na cadeia evolutiva dos dinossauros, o Santanaraptor ocuparia uma posição no grupo Tyrannoraptora, o mesmo do Tyrannossaurus rex, que habitou os EUA no final da era dos dinos.

Segundo Kellner, apesar de o animal ser um baixinho (poderia atingir, no máximo, 2,5 metros de altura), suas patas e bacia têm características anatômicas muito semelhantes às do ilustre réptil norte-americano.

"O Santanaraptor pode ser a espécie que deu origem ao tiranossauro 68 milhões de anos mais tarde", explicou o geólogo.

O exemplar de Santanaraptor encontrado pela equipe carioca foi desenterrado em 1991, mas a montagem do fóssil só foi concluída nove anos mais tarde. Tudo o que sobrou dele foram as patas e partes da cauda e da bacia, mas os pesquisadores conseguiram estimar que o bicho fosse um filhote de 1,5 metro de altura.

Sua estrutura óssea é de um dinossauro ágil e veloz, que provavelmente se alimentava de pequenas presas _um raptor, na linguagem dos paleontólogos. O nome é uma alusão à região onde ele viveu (a Formação Santana).

Texto marcado pelo PALAVRAS:

Pesquisadores=do=Museu=Nacional=do=Rio=de=Janeiro
[Pesquisadores=do=Museu=Nacional=do=Rio=de=Janeiro] <hum> PROP M P @SUBJ>
anunciaram [anunciar] <fmc> <mv> V PS/MQP 3P IND VFIN @FS-STA
ontem [ontem] ADV @<ADVL
a [o] <artd> DET F S @>N
descoberta [descoberta] <event> N F S @<ACC
de [de] <np-close> PRP @N<
uma [um] <arti> DET F S @>N
nova [novo] ADJ F S @>N
espécie [espécie] <A> N F S @P<
de [de] <np-close> PRP @N<
dinossauro [dinossauro] <Azo> N M S @P<
em [em] <sam-> PRP @<ADVL
o [o] <artd> <-sam> DET M S @>N
Brasil [Brasil] <civ> PROP M S @P<

$.
</s>
O [o] <artd> DET M S @>N
animal [animal] <Azo> N M S @SUBJ>
era [ser] <fmc> <mv> V IMPF 3S IND VFIN @FS-STA
um [um] <arti> DET M S @>N
carnívoro [carnívoro] <n> ADJ M S @<SC
que [que] <clb> KS @SUB
habitou [habitar] <fmc> <mv> V PS 3S IND VFIN @FS-STA
o [o] <artd> DET M S @>N
nordeste [nordeste] <dir> N M S @<ACC
brasileiro [brasileiro] <nat> <np-close> ADJ M S @N<
há [haver] <fmc> <mv> V PR 3S IND VFIN @FS-STA
110 [110] <NER:quantity> <card> NUM M P @>N
milhões [milhão] <NER2> <amount> N M P @<ACC
de [de] <np-close> PRP @N<
anos [ano] <dur> <per> N M P @P<
$,
em [em] <sam-> PRP @<ADVL
o [o] <artd> <-sam> DET M S @>N
período [período] <per> N M S @P<
Cretáceo [cretáceo] <prop> <np-close> ADJ M S @N<
$(
o [o] <artd> DET M S @>N
último [último] <n> <NUM-ord> ADJ M S @APP
de [de] <sam-> <np-close> PRP @N<
a [o] <artd> <-sam> DET F S @>N
era [era] <per> N F S @P<
de [de] <sam-> <np-close> PRP @N<
os [o] <artd> <-sam> DET M P @>N
grandes [grande] ADJ M P @>N
répteis [réptil] <Azo> N M P @P<
$)
$.
</s>
Batizado=de=Santanaraptor [Batizado=de=Santanaraptor] <common> PROP @SUBJ>
placidus [placer] <np-close> V PCP M P @N<
$, [$,] <co-subj> PU @CO
o [o] <artd> DET M S @>N
fóssil [fóssil] <A> N M S @SUBJ>
é [ser] <fmc> <mv> V PR 3S IND VFIN @FS-STA
o [o] <artd> DET M S @>N
único [único] <n> ADJ M S @<SC
a [a] PRP @<ADVL
ser [ser] <aux> V INF @ICL-P<
encontrado [encontrar] <mv> V PCP M S @ICL-AUX<
em [em] <sam-> PRP @<ADVL
o [o] <artd> <-sam> DET M S @>N
país [país] <Lciv> N M S @P<

com [com] PRP @<ADVL
restos [resto] <amount> N M P @P<
de [de] <np-close> PRP @N<
tecido [tecido] <mat-cloth> N M S @P<
mole [mole] <np-close> ADJ M S @N<
$,
como [como] <prd> <adv-rel> PRP @<SC
fibras [fibra] <asarg> <cord> N F P @P<
musculares [muscular] <np-close> ADJ F P @N<
$,
vasos [vaso] <asarg> <con> N M P @P<
sanguíneos [sanguíneo] <np-close> ADJ M P @N<
e [e] KC @CO
pele [pele] <asarg> <anbo> <mat-cloth> N F S @P<
$.
</s>
Para [para] PRP @ADVL>
os [o] <artd> DET M P @>N
paleontólogos [paleontólogo] <Hprof> N M P @P<
$,
achar [achar] <mv> V INF @IMV
esse [esse] <dem> DET M S @>N
tipo [tipo] <meta> N M S @SUBJ>
de [de] <np-close> PRP @N<
evidência [evidência] <percep-f> N F S @P<
equivale [equivaler] <fmc> <mv> V PR 3S IND VFIN @FS-STA
a [a] PRP @<PIV
acertar [acertar] <mv> V INF @ICL-P<
em [em] <sam-> PRP @<ADVL
a [o] <artd> <-sam> DET F S @>N
loteria [loteria] <occ> N F S @P<
$.
</s>
$"
É [ser] <mv> V PR 3S IND VFIN @FS-ACC>>
como=se [como=se] <clb> KS @SUB
o [o] <artd> DET M S @>N
dinossauro [dinossauro] <Azo> N M S @SUBJ>
tivesse [ter] <aux> V IMPF 3S SUBJ VFIN @FS-<SC
sido [ser] <aux> V PCP @ICL-AUX<
enterrado [enterrar] <mv> V PCP M S @ICL-AUX<
ontem [ontem] ADV @<ADVL
$"
$,
disse [dizer] <fmc> <mv> V PS 3S IND VFIN @FS-STA
Alexander=Kellner [Alexander=Kellner] <hum> PROP M S @<SUBJ
$,
geólogo [geólogo] <Hprof> <np-close> N M S @N<PRED
de [de] <sam-> <np-close> PRP @N<
o [o] <artd> <-sam> DET M S @>N

Setor=de=Paleovertebrados [Setor=de=Paleovertebrados] <inst> PROP M S @P<
de [de] <sam-> <np-close> PRP @N<
o [o] <artd> <-sam> DET M S @>N
Museu=Nacional [Museu=Nacional] <inst> PROP M S @P<
e [e] <co-prparg> KC @CO
coordenador [coordenador] <n> ADJ M S @P<
de [de] <sam-> <np-close> PRP @N<
a [o] <artd> <-sam> DET F S @>N
expedição [expedição] <occ> <HH> <act> N F S @P<
que [que] <clb> <rel> SPEC M S @SUBJ>
encontrou [encontrar] <mv> <np-close> V PS 3S IND VFIN @FS-N<
o [o] <artd> DET M S @>N
fóssil [fóssil] <A> N M S @<ACC
em [em] <sam-> <np-close> PRP @N<
a [o] <artd> <-sam> DET F S @>N
região [região] <Ltop> N F S @P<
de [de] <sam-> <np-close> PRP @N<
a [o] <artd> <-sam> DET F S @>N
Chapada=do=Araripe [Chapada=do=Araripe] <top> PROP F S @P<
$,
Ceará [Ceará] <hum> PROP M S @SUBJ>
$(
veja [ver] <fmc> <mv> V PR 3S SUBJ VFIN @FS-STA
mapa [mapa] <cc-r> N M S @<ACC
$)
$.
</s>
Com [com] PRP @PRED>
os [o] <artd> DET M P @>N
tecidos [tecido] <cc-rag> N M P @P<
preservados [preservar] <np-close> V PCP M P @N<PRED
$,
os [o] <artd> DET M P @>N
cientistas [cientista] <Hprof> N M P @SUBJ>
esperam [esperar] <fmc> <mv> V PR 3P IND VFIN @FS-STA
poder [poder] <aux> V INF @ICL-<ACC
saber [saber] <mv> V INF @ICL-AUX<
mais [muito] <KOMP> <quant> DET M/F S/P @<ACC
sobre [sobre] PRP @<ADVL
o [o] <artd> DET M S @>N
modo=de=vida [modo=de=vida] <f-psych> N M S @P<
e [e] <co-prparg> <co-prparg> KC @CO
a [o] <artd> DET F S @>N
evolução [evolução] <process> N F S @P<
de [de] <sam-> <np-close> PRP @N<
os [o] <artd> <-sam> DET M P @>N
répteis [réptil] <Azo> N M P @P<
$.
</s>
Outra [outro] <KOMP> <diff> DET F S @>N

```
importante	[importante] ADJ F S @>N
descoberta	[descoberta] <event> N F S @SUBJ>
é=que	[é=que] <foc> ADV @<FOC
$,
em	[em] <sam-> PRP @ADVL>
a	[o] <artd> <-sam> DET F S @>N
cadeia	[cadeia] <inst> N F S @P<
evolutiva	[evolutivo] <np-close> ADJ F S @N<
de	[de] <sam-> <np-close> PRP @N<
os	[o] <artd> <-sam> DET M P @>N
dinossauros	[dinossauro] <Azo> N M P @P<
$,	[$,] <co-subj> PU @CO
o	[o] <artd> DET M S @>N
Santanaraptor	[Santanaraptor] <inst> PROP M S @SUBJ>
ocuparia	[ocupar] <fmc> <mv> V COND 3S VFIN @FS-STA
uma	[um] <arti> DET F S @>N
posição	[posição] <pos-an> <act> N F S @<ACC
em	[em] <sam-> PRP @<ADVL
o	[o] <artd> <-sam> DET M S @>N
grupo	[grupo] <HH> N M S @P<
Tyrannoraptora	[Tyrannoraptora] <org> <np-close> PROP M S @N< @N<
$,	[$,] <co-acc> PU @CO
o	[o] <artd> DET M S @>N
mesmo	[mesmo] <diff> <KOMP> DET M S @<ACC
de	[de] <sam-> PRP @<ADVL
o	[o] <artd> <-sam> DET M S @>N
Tyrannossaurus=rex	[Tyrannossaurus=rex] <inst> PROP M S @P< @P<
$,
que	[que] <clb> <rel> SPEC M S @SUBJ>
habitou	[habitar] <mv> <np-close> V PS 3S IND VFIN @FS-N<
os	[o] <artd> DET M P @>N
EUA	[EUA] <civ> PROP M P @<ACC
em	[em] <sam-> PRP @<ADVL
o	[o] <artd> <-sam> DET M S @>N
final	[final] <event> N M S @P<
de	[de] <sam-> <np-close> PRP @N<
a	[o] <artd> <-sam> DET F S @>N
era	[era] <per> N F S @P<
de	[de] <sam-> <np-close> PRP @N<
os	[o] <artd> <-sam> DET M P @>N
dinos	[dino] <Azo> N M P @P<
$.
</s>
Segundo	[segundo] <com> <adv-rel> PRP @ADVL>
Kellner	[Kellner] <hum> <asarg> PROP M/F S @P<
$,
apesar=de	[apesar=de] PRP @ADVL>
o	[o] <clb> <artd> DET M S @>N
animal	[animal] <Azo> N M S @SUBJ>
ser	[ser] <mv> V INF 3S @ICL-P<
```

um [um] <arti> DET M S @>N
baixinho [baixinho] <n> ADJ M S @<SC
$(
poderia [poder] <fmc> <aux> V COND 3S VFIN @FS-STA
atingir [atingir] <mv> V INF @ICL-AUX<
$, [$,] <co-fmc> PU @CO
no=máximo [no=máximo] ADV @<ADVL
$, [$,] <co-fmc> PU @CO
2,5 [2,5] <NER:quantity> <card> NUM M P @>N
metros [metro] <NER2> <unit> N M P @<ACC
de [de] <np-close> PRP @N<
altura [altura] <f-q> N F S @P<
$)
$, [$,] <co-fmc> <co-fin> PU @CO
suas [seu] <poss 3S> DET F P @>N
patas [pata] <anzo> N F P @SUBJ>
e [e] <co-subj> <co-subj> KC @CO
bacia [bacia] <Ltop> <con> N F S @SUBJ>
têm [ter] <fmc> <mv> V PR 3P IND VFIN @FS-STA
características [característica] <ac> N F P @<ACC
anatômicas [anatômico] <np-close> ADJ F P @N<
muito [muito] <quant> ADV @>A
semelhantes [semelhante] <np-close> ADJ M/F P @N<
a [a] <sam-> PRP @A<
as [o] <dem> <-sam> DET F P @P<
de [de] <sam-> <np-close> PRP @N<
o [o] <artd> <-sam> DET M S @>N
ilustre [ilustre] ADJ M S @>N
réptil [réptil] <Azo> N M S @P<
norte-americano [norte-americano] <nat> <np-close> ADJ M S @N<
$.
</s>
$"
O [o] <artd> DET M S @>N
Santanaraptor [Santanaraptor] <inst> PROP M S @SUBJ>
pode [poder] <aux> V PR 3S IND VFIN @FS-ACC>>
ser [ser] <mv> V INF @ICL-AUX<
a [o] <artd> DET F S @>N
espécie [espécie] <meta> N F S @<SC
que [que] <clb> <rel> SPEC M S @SUBJ>
deu [dar] <mv> <np-close> V PS 3S IND VFIN @FS-N<
origem [origem] <Labs> N F S @<ACC
a [a] <sam-> PRP @<PIV
o [o] <-sam> <artd> DET M S @>N
tiranossauro [tiranossauro] <Adom> N M S @P<
68 [68] <NER:quantity> <card> NUM M P @>N
milhões [milhão] <NER2> <amount> N M P @PRED>
de [de] <np-close> PRP @N<
anos [ano] <dur> <per> N M P @P<
mais [mais] ADV @ADVL>

tarde [tarde] <pp> ADV @<ADVL
$"
$,
explicou [explicar] <fmc> <mv> V PS 3S IND VFIN @FS-STA
o [o] <artd> DET M S @>N
geólogo [geólogo] <Hprof> N M S @<SUBJ
$.
</s>
O [o] <artd> DET M S @>N
exemplar [exemplar] <cc> N M S @SUBJ>
de [de] <np-close> PRP @N<
Santanaraptor [Santanaraptor] <hum> PROP M/F S @P<
encontrado [encontrar] <mv> <np-close> V PCP M S @ICL-N<
por [por] <sam-> PRP @<PASS
a [o] <artd> <-sam> DET F S @>N
equipe [equipe] <HH> N F S @P<
carioca [carioca] <nat> <np-close> ADJ F S @N<
foi [ser] <fmc> <aux> V PS 3S IND VFIN @FS-STA
desenterrado [desenterrar] <mv> V PCP M S @ICL-AUX<
em [em] PRP @<ADVL
1991 [1991] <date> <card> <NER:date> NUM M/F P @P<
$,
mas [mas] <co-fin> <co-fmc> <co-fin> KC @CO
a [o] <artd> DET F S @>N
montagem [montagem] <act> N F S @SUBJ>
de [de] <sam-> <np-close> PRP @N<
o [o] <artd> <-sam> DET M S @>N
fóssil [fóssil] <A> N M S @P<
só [só] ADV @ADVL>
foi [ser] <fmc> <aux> V PS 3S IND VFIN @FS-STA
concluída [concluir] <mv> V PCP F S @ICL-AUX<
nove [nove] <card> NUM M P @>N
anos [ano] <dur> N M P @<ADVL
mais [mais] <np-close> ADV @N<
tarde [tarde] <pp> ADV @<ADVL
$.
</s>
Tudo=o=que [tudo=o=que] <clb> <quant> <rel> SPEC M S @SUBJ>
sobrou [sobrar] <mv> V PS 3S IND VFIN @FS-SUBJ>
de [de] <sam-> PRP @<ADVL
ele [ele] <-sam> PERS M 3S NOM/PIV @P<
foram [ser] <fmc> <mv> V PS/MQP 3P IND VFIN @FS-STA
as [o] <artd> DET F P @>N
patas [pata] <anzo> N F P @<SC
e [e] <co-fin> <co-fmc> KC @CO
partes [parte] <HH> N F P @<ACC
de [de] <sam-> <np-close> PRP @N<
a [o] <artd> <-sam> DET F S @>N
cauda [cauda] <anmov> N F S @P<
e [e] <co-postnom> <co-sc> KC @CO

de [de] <sam-> PRP @<SC
a [o] <artd> <-sam> DET F S @>N
bacia [bacia] <Ltop> <con> N F S @P<
$, [$,] <co-subj> <co-fin> PU @CO
mas [mas] <co-subj> <co-fin> KC @CO
os [o] <artd> DET M P @>N
pesquisadores [pesquisador] <Hprof> N M P @SUBJ>
conseguiram [conseguir] <fmc> <mv> V PS/MQP 3P IND VFIN @FS-STA
estimar [estimar] <mv> V INF @ICL-<ACC
que [que] <clb> KS @SUB
o [o] <artd> DET M S @>N
bicho [bicho] <H> <Azo> N M S @SUBJ>
fosse [ser] <mv> V IMPF 3S SUBJ VFIN @FS-<ACC
um [um] <arti> DET M S @>N
filhote [filhote] <H> N M S @<SC
de [de] <np-close> PRP @N<
1,5 [1,5] <NER:quantity> <card> NUM M P @>N
metro [metro] <NER2> <unit> N M S @P<
de [de] <np-close> PRP @N<
altura [altura] <f-q> N F S @P<
$.
</s>
Sua [seu] <poss 3S> DET F S @>N
estrutura [estrutura] <percep-f> N F S @SUBJ>
óssea [ósseo] <np-close> ADJ F S @N<
é [ser] <fmc> <mv> V PR 3S IND VFIN @FS-STA
de [de] PRP @<SC
um [um] <arti> DET M S @>N
dinossauro [dinossauro] <Azo> N M S @P<
ágil [ágil] ADJ M/F S @<PRED
e [e] <co-pred> KC @CO
veloz [veloz] ADJ M/F S @<PRED
$,
que [que] <clb> <rel> SPEC M S @SUBJ>
provavelmente [provavelmente] ADV @ADVL>
se [se] PERS M/F 3S ACC @ACC>-PASS
alimentava [alimentar] <mv> <np-close> V IMPF 3S IND VFIN @FS-N<
de [de] PRP @<PIV
pequenas [pequeno] ADJ F P @>N
presas [presa] <anzo> <A> <cc> N F P @P<
$_
um [um] <arti> DET M S @>N
raptor [raptor] <H> N M S @PRED>
$,
em [em] <sam-> PRP @<ADVL
a [o] <artd> <-sam> DET F S @>N
linguagem [linguagem] <sem-s> N F S @P<
de [de] <sam-> <np-close> PRP @N<
os [o] <artd> <-sam> DET M P @>N
paleontólogos [paleontólogo] <Hprof> N M P @P<

$.
</s>
O [o] <artd> DET M S @>N
nome [nome] <percep-f> N M S @SUBJ>
é [ser] <fmc> <mv> V PR 3S IND VFIN @FS-STA
uma [um] <arti> DET F S @>N
alusão [alusão] <act-s> N F S @<SC
a [a] <sam-> <np-close> PRP @N<
a [o] <-sam> <artd> DET F S @>N
região [região] <Ltop> <L> N F S @P<
onde [onde] <clb> <ks> <rel> ADV @ADVL>
ele [ele] PERS M 3S NOM @SUBJ>
viveu [viver] <mv> <np-close> V PS 3S IND VFIN @FS-N<
$(
a [o] <artd> DET F S @>N
Formação=Santana [Formação=Santana] <occ> PROP F S @APP
$)
$.

Texto gerado pelo parser secundário com marcas de interesse:

PN|SBJ|Pesquisadores do Museu Nacional do Rio de Janeiro# Verbfsta|_|anunciaram|[anunciar]# ADV|ADL|ontem# DET|_|a# NOUM|ACC|descoberta# PRP|_|de# DET|_|uma# ADJ|_|nova# NOUM|_|espécie# PRP|_|de# NOUM|_|dinossauro# PRP|_|no# PN|_|Brasil# _|_|.# end|_|

DET|_|O# NOUM|SBJ|animal# Verbfsta|_|era|[ser]# DET|_|um# ADJ|_|carnívoro# KS|CS|que# Verbfsta|_|habitou|[habitar]# DET|_|o# NOUM|ACC|nordeste# ADJ|_|brasileiro# Verbfsta|_|há|[haver]# NUM|_|110# NOUM|ACC|milhões# PRP|_|de# NOUM|_|anos# _|_|,# PRP|_|no# NOUM|_|período# ADJ|_|Cretáceo# _|_|(# DET|_|o# ADJ|APP|último# PRP|_|da# NOUM|_|era# PRP|_|dos# ADJ|_|grandes# NOUM|_|répteis# _|_|)# _|_|.# end|_|

PN|SBJ|Batizado de Santanaraptor# Verbpcp|_|placidus|[placer]# _|CO|,# DET|_|o# NOUM|SBJ|fóssil# Verbfsta|_|é|[ser]# DET|_|o# ADJ|_|único# PRP|ADL|a# Verb|_|ser|[ser]# Verbpcp|_|encontrado|[encontrar]# PRP|_|no# NOUM|_|país# PRP|ADL|com# NOUM|_|restos# PRP|_|de# NOUM|_|tecido# ADJ|_|mole# _|_|,# PRP|_|como# NOUM|_|fibras# ADJ|_|musculares# _|_|,# NOUM|_|vasos# ADJ|_|sanguíneos# KC|CO|e# NOUM|_|pele# _|_|.# end|_|

PRP|ADL|Para# DET|_|os# NOUM|_|paleontólogos# _|_|,# Verb|_|achar|[achar]# DET|_|esse# NOUM|SBJ|tipo# PRP|_|de# NOUM|_|evidência# Verbfsta|_|equivale|[equivaler]# PRP|_|a# Verb|_|acertar|[acertar]# PRP|_|na# NOUM|_|loteria# _|_|.# end|_|

_|_|"# Verb|ACC|É|[ser]# KS|CS|como se# DET|_|o# NOUM|SBJ|dinossauro# Verb|_|tivesse|[ter]# Verbpcp|_|sido|[ser]# Verbpcp|_|enterrado|[enterrar]# ADV|ADL|ontem# _|_|"# _|_|,# Verbfsta|_|disse|[dizer]# PN|_|Alexander Kellner# _|_|,# NOUM|_|geólogo# PRP|_|do# PN|_|Setor de Paleovertebrados# PRP|_|do# PN|_|Museu Nacional# KC|CO|e# ADJ|_|coordenador# PRP|_|da# NOUM|_|expedição# SPECrel|SBJ|que# Verb|_|encontrou|[encontrar]# DET|_|o# NOUM|ACC|fóssil# PRP|_|na# NOUM|_|região# PRP|_|da# PN|_|Chapada do Araripe# _|_|,# PN|SBJ|Ceará# _|_|(# Verbfsta|_|veja|[ver]# NOUM|ACC|mapa# _|_|)# _|_|.# end|_|

PRP|_|Com# DET|_|os# NOUM|_|tecidos# Verbpcp|_|preservados|[preservar]# _|_|,# DET|_|os# NOUM|SBJ|cientistas# Verbfsta|_|esperam|[esperar]# Verb|ACC|poder|[poder]#

Verb|_|saber|[saber]# DET|ACC|mais# PRP|ADL|sobre# DET|_|o# NOUM|_|modo de vida# KC|CO|e# DET|_|a# NOUM|_|evolução# PRP|_|dos# NOUM|_|répteis# _|_|.# end|_|

DET|_|Outra# ADJ|_|importante# NOUM|SBJ|descoberta# ADV|_|é que# _|_|,# PRP|_|na# NOUM|_|cadeia# ADJ|_|evolutiva# PRP|_|dos# NOUM|_|dinossauros# _|CO|,# DET|_|o# PN|SBJ|Santanaraptor# Verbfsta|_|ocuparia|[ocupar]# DET|_|uma# NOUM|ACC|posição# PRP|_|no# NOUM|_|grupo# PN|_|Tyrannoraptora# _|CO|,# DET|_|o# DET|ACC|mesmo# PRP|_|do# PN|_|Tyrannossaurus rex# _|_|,# SPECrel|SBJ|que# Verb|_|habitou|[habitar]# DET|_|os# PN|ACC|EUA# PRP|_|no# NOUM|_|final# PRP|_|da# NOUM|_|era# PRP|_|dos# NOUM|_|dinos# _|_|.# end|_|

PRP|ADLrel|Segundo# PN|_|Kellner# _|_|,# PRP|ADL|apesar de# DET|_|o# NOUM|SBJ|animal# Verb|_|ser|[ser]# DET|_|um# ADJ|_|baixinho# _|_|(# Verbfsta|_|poderia|[poder]# Verb|_|atingir|[atingir]# _|COfmc|,# ADV|ADL|no máximo# _|COfmc|,# NUM|_|2,5# NOUM|ACC|metros# PRP|_|de# NOUM|_|altura# _|_|)# _|COfmc|,# PRO|_|suas# NOUM|SBJ|patas# KC|CO|e# NOUM|SBJ|bacia# Verbfsta|_|têm|[ter]# NOUM|ACC|características# ADJ|_|anatômicas# ADV|_|muito# ADJ|_|semelhantes# PRP|_|a# PRP|_|do# ADJ|_|ilustre# NOUM|_|réptil# ADJ|_|norte-americano# _|_|.# end|_|

_|_|"# DET|_|O# PN|SBJ|Santanaraptor# Verb|ACC|pode|[poder]# Verb|_|ser|[ser]# DET|_|a# NOUM|_|espécie# SPECrel|SBJ|que# Verb|_|deu|[dar]# NOUM|ACC|origem# PRP|_|ao# NOUM|_|tiranossauro# NUM|_|68# NOUM|_|milhões# PRP|_|de# NOUM|_|anos# ADV|ADL|mais# ADV|ADL|tarde# _|_|"# _|_|,# Verbfsta|_|explicou|[explicar]# DET|_|o# NOUM|_|geólogo# _|_|.# end|_|

DET|_|O# NOUM|SBJ|exemplar# PRP|_|de# PN|_|Santanaraptor# Verbpcp|_|encontrado|[encontrar]# PRP|_|pela# NOUM|_|equipe# ADJ|_|carioca# Verbfsta|_|foi|[ser]# Verbpcp|_|desenterrado|[desenterrar]# PRP|ADL|em# NUM|_|1991# _|_|,# KC|COfmc|mas# DET|_|a# NOUM|SBJ|montagem# PRP|_|do# NOUM|_|fóssil# ADV|ADL|só# Verbfsta|_|foi|[ser]# Verbpcp|_|concluída|[concluir]# NUM|_|nove# NOUM|ADL|anos# ADV|_|mais# ADV|ADL|tarde# _|_|.# end|_|

SPECrel|SBJ|Tudo o que# Verb|_|sobrou|[sobrar]# PRP|_|dele# Verbfsta|_|foram|[ser]# DET|_|as# NOUM|_|patas# KC|COfmc|e# NOUM|ACC|partes# PRP|_|da# NOUM|_|cauda# KC|CO|e# PRP|_|da# NOUM|_|bacia# _|COfin|,# KC|COfin|mas# DET|_|os# NOUM|SBJ|pesquisadores# Verbfsta|_|conseguiram|[conseguir]# Verb|ACC|estimar|[estimar]# KS|CS|que# DET|_|o# NOUM|SBJ|bicho# Verb|ACC|fosse|[ser]# DET|_|um# NOUM|_|filhote# PRP|_|de# NUM|_|1,5# NOUM|_|metro# PRP|_|de# NOUM|_|altura# _|_|.# end|_|

PRO|_|Sua# NOUM|SBJ|estrutura# ADJ|_|óssea# Verbfsta|_|é|[ser]# PRP|_|de# DET|_|um# NOUM|_|dinossauro# ADJ|PRED|ágil# KC|CO|e# ADJ|PRED|veloz# _|_|,# SPECrel|SBJ|que# ADV|ADL|provavelmente# PRO|ACC|se# Verb|_|alimentava|[alimentar]# PRP|_|de# ADJ|_|pequenas# NOUM|_|presas# _|_|_# DET|_|um# NOUM|_|raptor# _|_|,# PRP|_|na# NOUM|_|linguagem# PRP|_|dos# NOUM|_|paleontólogos# _|_|.# end|_|

DET|_|O# NOUM|SBJ|nome# Verbfsta|_|é|[ser]# DET|_|uma# NOUM|_|alusão# PRP|_|à# NOUM|_|região# ADVrel|ADLkc|onde# PRO|SBJ|ele# Verb|_|viveu|[viver]# _|_|(# DET|_|a# PN|APP|Formação Santana# _|_|)# _|_|.#

Texto segmentado, com marcas de relações complementares e com linhas numeradas:

1: Pesquisadores do Museu Nacional do Rio de Janeiro anunciaram ontem a descoberta de uma nova espécie de dinossauro no Brasil .

2: O animal era um carnívoro que habitou o nordeste brasileiro há 110 milhões de anos , no período Cretáceo

3: {PAR} (o último da era dos grandes répteis).

4: Batizado de Santanaraptor placidus , o fóssil é o único a ser encontrado no país com restos de tecido mole , como fibras musculares , vasos sanguíneos e pele .

5: Para os paleontólogos , achar esse tipo de evidência equivale a acertar na loteria .

6: " É como se o dinossauro tivesse sido enterrado ontem " ,

7: {ATB} disse Alexander Kellner , geólogo do Setor de Paleovertebrados do Museu Nacional e coordenador da expedição

8: que encontrou o fóssil na região da Chapada do Araripe , Ceará

9: {PAR} (veja mapa).

10: Com os tecidos preservados , os cientistas esperam poder saber mais sobre o modo de vida e a evolução dos répteis .

11: Outra importante descoberta é que , na cadeia evolutiva dos dinossauros , o Santanaraptor ocuparia uma posição no grupo Tyrannoraptora , o mesmo do Tyrannossaurus rex ,

12: que habitou os EUA no final da era dos dinos .

13: {SUT1} Segundo Kellner ,apesar de o animal ser um baixinho

14: {PAR} (poderia atingir , no máximo , 2,5 metros de altura)

15: {SUT2} , --suas patas e bacia têm características anatômicas muito semelhantes a do ilustre réptil norte-americano . " O Santanaraptor pode ser a espécie

16: que deu origem ao tiranossauro 68 milhões de anos mais tarde " ,

17: explicou o geólogo .

18: O exemplar de Santanaraptor encontrado pela equipe carioca foi desenterrado em 1991 ,

19: mas --a montagem do fóssil só foi concluída nove anos mais tarde .

20: Tudo o que sobrou dele foram as patas e partes da cauda e da bacia

21: , --mas --os pesquisadores conseguiram estimar que o bicho fosse um filhote de 1,5 metro de altura .

22: Sua estrutura óssea é de um dinossauro ágil e veloz ,

23: que provavelmente se alimentava de pequenas presas

24: {PAR} _um raptor , na linguagem dos paleontólogos .

25: O nome é uma alusão à região

26: onde ele viveu

27: (a Formação Santana).

Apêndice B: detalhes de execução

O método em foco foi desenvolvido em Perl, uma linguagem de programação famosa por seus recursos de tratamento de textos, dos quais se destacam as expressões regulares, que são padrões para busca e substituição em textos através de regras (expressões) bem definidas.

Os passos de execução deste método são representados por programas em linguagem Perl. O primeiro passo é o processamento do texto original através do parser PALAVRAS, feito chamando o parser e passando para ele o arquivo com o texto de entrada. A saída do parser é gravada em outro arquivo, contendo toda a marcação. Este último é usado por outro programa que busca apenas as marcas de interesse para o processo de segmentação, além de juntar as preposições e artigos que foram separados pelo parser, tais como "do" que se tornou "de" e "o", decomposição necessária a uma análise morfossintática. Este programa, chamado de parser secundário, coloca duas marcações para cada palavra: a primeira é sua classe gramatical e, a segunda, sua função sintática, sendo que essas marcas ficam vazias em palavras que não têm função de interesse.

O arquivo gerado com as marcas de interesse e contendo uma sentença por linha é processado pelo programa que contém as regras de segmentação na forma de expressões regulares; cada sentença é processada palavra a palavra, levando em consideração, na maioria das regras, a palavra anterior e a posterior à palavra em análise.

Como os segmentos gerados podem não conter um verbo, excetuando-se o caso de UBDs acopladas, um outro programa faz a verificação de cada segmento gerado pelo programa anterior, juntado ao segmento que não contém um verbo ao próximo segmento. Isto leva a melhores resultados, dado que as regras de segmentação criam, em alguns casos, segmentos que não equivalem a uma UBD.

Para a colocação das marcas que auxiliarão o analisador retórico automático, outro programa varre os segmentos em busca de UBDs acopladas e verifica se estas ocorrem no interior de uma UBD; se assim for, marcam-se as partes do segmento partido, indicando que estas partes formam um mesmo segmento (de acordo com a relação *same-unit* da RST).

Por fim um programa final numera os segmentos sequencialmente, visto que estão cada um em uma linha e prontos para a análise retórica, tanto automática quanto manual.