A Floresta Sintá(c)tica como recurso para estudar a língua portuguesa

Diana Santos, Eckhard Bick, Raquel Marchi e Susana Afonso

Quando se aprende ou ensina gramática, as frases usadas são normalmente muito simples – demasiado simples, de fato, para ilustrarem a forma como a língua é usada.

Uma floresta sintática (do inglês "treebank") é uma coleção de frases reais, que foram produzidas num dado contexto comunicativo, analisadas lingüisticamente. Ou seja, é um banco de dados sobre a língua, contendo texto produzido em outro contexto que não o de ensinar ou ilustrar a sintaxe do português.

No nosso projeto, usamos texto publicado em jornais – a *Folha de São Paulo* e o *Público* – para criar o primeiro recurso desse tipo para a língua portuguesa. Esse texto foi, primeiro, analisado automaticamente pelo PALAVRAS (Bick, 2000) e, depois, revisto por uma equipa de lingüistas.

Introdução

O projeto Floresta Sintá(c)tica (http://www.linguateca.pt/Floresta/) é uma colaboração entre duas iniciativas independentes, o projeto VISL e a Linguateca, e surgiu como continuação ou extensão do projeto AC/DC (Santos & Bick, 2000; Santos & Sarmento, 2003), que dá acesso na Rede (Web) a grande quantidade de texto automaticamente anotado pelo analisador sintático PALAVRAS.

O material servido pelo AC/DC tem, contudo, duas limitações importantes: o fato de as análises não serem revisadas e confirmadas – ou corrigidas – por seres humanos; e o fato da marcação produzida pelo PALAVRAS ser de gramática dependencial, não marcando, portanto, explicitamente constituintes, nem decidindo qual a ligação (em inglês,

“attachment”) entre eles. Por exemplo, no texto *“o carro de linhas da Maria*”, o analisador sintático indicaria que o sintagma preposicional *“da Maria*” estava ligado a um sintagma nominal à sua esquerda, mas não especificaria a qual: a *“linhas*”, ou a “*carro de linhas*”? O projeto Floresta Sintá(c)tica foi lançado precisamente para adicionar este tipo de informação, e rever a anotação automática, que podia ter sido errônea, ou simplesmente faltar.

Afonso et al. (2002a) descreve os primeiros passos do projeto – veja-se também (Afonso et al. 2002b; 2002c), duas versões mais reduzidas do mesmo documento. Dois anos e meio mais tarde, existe documentação mais estruturada e completa, em constante desenvolvimento (Afonso, 2004a). De momento, temos disponíveis – quer para consulta quer para obtenção na sua totalidade, formando aquilo a que chamamos o *Bosque* – cerca de 7500 árvores revistas, correspondendo a aproximadamente 150 mil palavras, e este número vai aumentando com o desenrolar do projeto. O Bosque existe em vários formatos, em particular um dependencial e outro sintagmático (com a estrutura de constituintes explícita).

Na figura 1, apresentamos a saída do analisador sintático e sua revisão, ainda em formato dependencial, para a frase (um título do Público) *“Acontece, na TV2, com Zombi dos Palmares*”, frase essa que, para a sua análise correta, exige saber que existe um programa de televisão chamado “*Acontece*” e que “*Zombi dos Palmares*” é um nome próprio.

Análise automática		Revisão humana	
Acontece	**[acontecer] <fmc> V PR 3S IND VFIN @FMV**	**Acontece**	**[Acontece] PROP M S @NPHR**
$,		$,	
em	[em] <sam-> PRP **@<ADVL**	em	[em] <sam-> PRP **@N<PRED**
a	[a] <-sam> <artd> DET F S @>N	a	[o] <-sam> <artd> DET F S @>N
TV2	[TV2] PROP F S @P<	TV2	[TV2] PROP F S @P<
$,		$,	
com	[com] PRP @N<	com	[com] PRP @N<
Zombi	**[Zombi] PROP M/F S @P<**	**Zombi=dos=Palmares**	**[Zombi=dos=Palmares] PROP M S @P<**
de	**[de] <sam-> PRP @N<**		
os	**[o] <-sam> <artd> DET M P @>N**		
Palmares	**[palmar] <prop> N M P @P<**		

Figura1. Exemplo de formato dependencial antes e depois da revisão humana

Em seguida, a sua codificação em formato de constituintes (por nós chamado de "árvores deitadas") é obtida automaticamente e revisada, dando origem ao seguinte:

```
CP641-1 Acontece, na TV2, com Zombi dos Palmares
A1
UTT:np
=H:prop('Acontece'  M S)            Acontece
=,
=N<PRED:pp
==H:prp('em' <sam->)      em
==P<:np
===>N:art('o' <-sam> <artd> F S)  a
===H:prop('TV2' F S)      TV2
=,
=N<:pp
==H:prp('com')   com
==P<:prop('Zombi_dos_Palmares' M S)
Zombi_dos_Palmares
```

Figura 2. Formato árvores deitadas, revisado

Motivação

Houve várias razões distintas, mas possivelmente complementares, para criar a Floresta. Em primeiro lugar, um recurso destes estava faltando para o português, embora já corrente para outras línguas (veja-se o Penn Treebank (Marcus et al., 1993) e o SUSANNE (Sampson, s/d) para o inglês, ou o Prague Dependency Treebank (Haji?, 1998) para o checo), e um dos objetivos da Linguateca (Santos, 2002) é criar recursos que facilitem o progresso do processamento computacional da nossa língua.

O VISL (http://visl.sdu.dk/), por seu lado, tem como principal objetivo criar materiais de ensino na Web. Ora a exposição a texto real, como o incorporado na Floresta, é indubitavelmente relevante por razões pedagógicas; além disso, a criação e revisão de um conjunto de frases anotadas pelo seu analisador sintático é uma das melhores formas, senão a melhor, de melhorar e tornar mais robusto o próprio PALAVRAS.

Finalmente, outra das motivações para a construção da Floresta Sintá(c)tica é o desejo de obter um consenso que possa depois ser usado na avaliação de outros analisadores automáticos, segundo o modelo de avaliação conjunta preconizado em Santos (no prelo).

Houve várias razões para a escolha inicial de material jornalístico: em primeiro lugar, devido à maior facilidade de obter o direito de o redistribuir livremente; em segundo lugar, porque se presume que a audiência típica de um jornal é maior do que a de obras literárias ou técnicas. Contudo, essa é uma opção que não é fundamental, e é possível que a Floresta venha mais tarde a incorporar outro gênero de texto.

Alguns exemplos de revisão

Num espaço tão limitado, é difícil dar uma panorâmica quer do tipo de decisões envolvidas quer das potencialidades de uso do material. Mencionamos, portanto, que existe documentação extensa sobre a Floresta no próprio site da rede, e que tentamos sempre responder rápida e adequadamente a qualquer pergunta dos usuários. Apresentaremos aqui apenas alguns exemplos simples do que está envolvido na revisão, deixando questões mais complexas para outros artigos.

Muitas palavras podem receber classificação diferente, dependendo do contexto em que foram inseridas em uma determinada sentença. Os exemplos mais comuns deste tipo de ambigüidade, são os de palavras que podem ser classificadas tanto como substantivo, quanto adjetivo (palavras como "*brasileiro/a*"; "*jovem*"), ou ainda, como adjetivos e advérbios ("*rápido/a*", por exemplo) e adjetivos e particípio passado de alguns verbos ("*passado/a*"; "*eleito/a*", etc.). É importante dizer que, às vezes, essa ambigüidade pode apresentar mais de uma leitura de um determinado sintagma ou sentença, e nestes casos a solução encontrada durante a revisão humana é a de trazer uma análise de acordo com cada leitura. Um exemplo

seria um sintagma nominal do tipo "*O jovem trabalhador*", que permite a leitura de "*jovem*" ora como modificador do núcleo do sintagma, ora como sendo o próprio núcleo, estando ambas corretas.

O contexto é fundamental para decidir o papel que cada palavra desempenha na sentença como um todo. Por exemplo, é geralmente preciso decidir se, em uma sentença, uma palavra ou grupo de palavras é o sujeito ou o objeto/predicativo do sujeito na oração. Muitas vezes, a ordem dos elementos na sentença pode confundir a análise automática, como foi o caso da sentença seguinte, "*Nova droga combate asma, diz estudo*", na figura 3, em que a palavra "*estudo*" é o núcleo do sintagma nominal sujeito da oração e não seu objeto direto, como havia sido anteriormente analisado.

```
STA:fcl
=ACC:fcl
==SUBJ:np
===>N:adj('novo' F S)     Nova
===H:n('droga' F S) droga
==P:v-fin('combater' PR 3S IND)  combate
==ACC:n('asma' F S) asma
=,
=P:v-fin('dizer' PR 3S IND)      diz
=SUBJ:n('estudo' M S)     estudo
```

Figura 3. Caso problemático para a análise automática

De forma resumida, as palavras podem ser ambíguas na morfologia, ou seja, na sua classificação individual (uma mesma palavra pode pertencer a classes gramaticais diferentes), ou na sintaxe, no papel ou função que ela desempenha dentro da sentença. Pois dependendo da sentença, até um simples "determinante" em um sintagma nominal, como o artigo, pode ocupar papel de núcleo deste. O exemplo seriam sentenças do tipo "*O garoto da direita havia feito a pergunta e o da esquerda devia respondê-la*". Na primeira oração, o sujeito "*o garoto da direita*" tem como núcleo a palavra "*garoto*". Na segunda oração, em vez, o núcleo do sujeito "*o da esquerda*" é "*o*", visto que a palavra "*garoto*" está implícita (elíptica) nesta oração. Neste caso, a palavra "*o*", artigo definido no masculino singular, vai

ter sempre a mesma classificação morfológica nas duas orações, mas assumirá diferentes papéis sintáticos (ora determinante, ora núcleo).

Nestes casos, a intervenção humana, ciente do contexto de cada palavra inserida na sentença, pode resolver facilmente aquilo que pode não ter sido resolvido no processamento automático. Sobre este assunto, consulte-se também Marchi (2004) e Afonso (2004b).

Alguns exemplos de uso

É possível, além de obter integralmente a Floresta em arquivo de texto simples em vários formatos, interagir com ela usando dois sistemas de procura diferentes, o CorpusEye (Bick, 2003; 2004) e o Águia (Santos, 2003).

Apresentamos alguns exemplos aqui, que não substituem de forma alguma o seu uso direto na Internet. A figura 4 mostra (um excerto d)o resultado do CorpusEye (http://corp.hum.sdu.dk) na extração de hierarquias de orações, ou seja, orações relativas contendo uma oração finita:

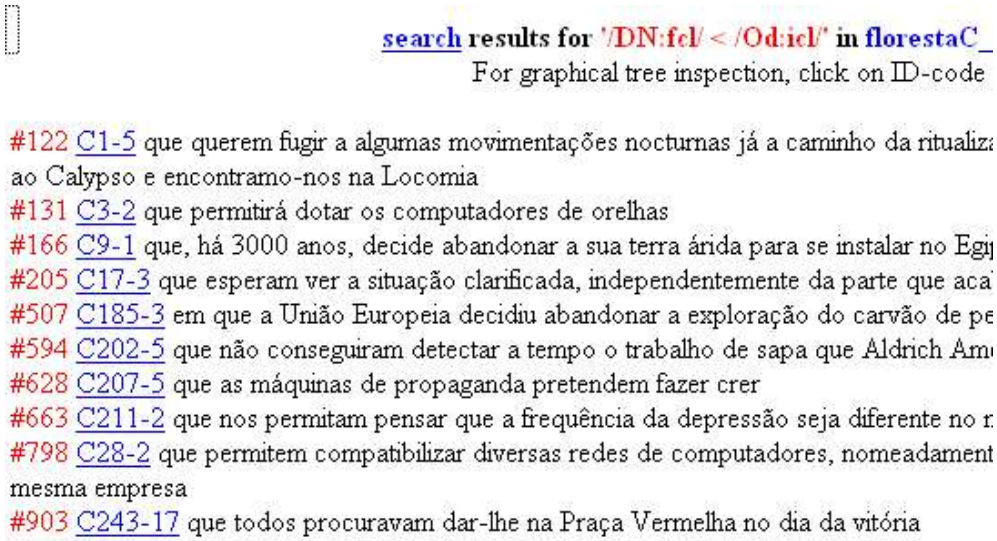

Figura 4. Extração, com o CorpusEye, de orações relativas contendo uma oração finita

Também é possível ver, e manipular, cada árvore no formato gráfico apresentado na figura 5. De fato, uma árvore sintática pode ser manipulada de vários modos: desdobrada

camada por camada, filtrada em termos de complexidade, ou até reconstruída de maneira interativa com objetivos didáticos.

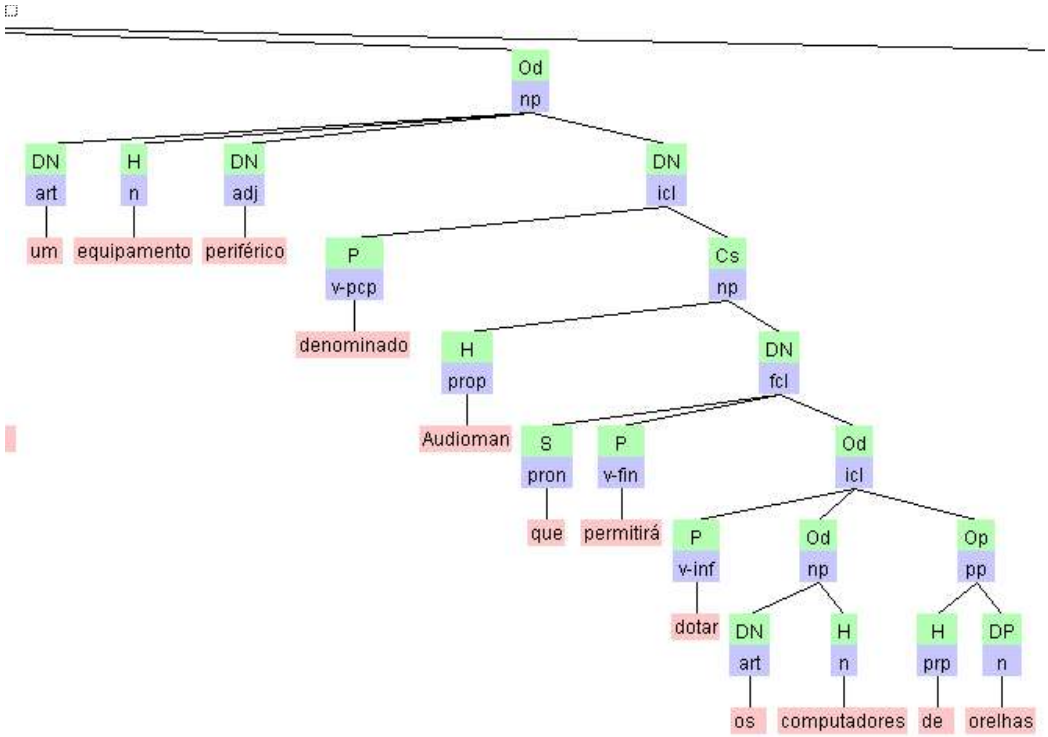

Figura 5. Uma árvore gráfica, manipulável com os programas do VISL

O CorpusEye também fornece uma interface baseada em menus, que permite procurar seqüências de categorias gramaticais, informação morfológica e função sintática, além de poder fornecer a distribuição dos resultados, normalizada em relação às freqüências do fenômeno em texto corrente.

Ilustrando agora as capacidades do Águia, apresentamos um excerto do seu resultado na figura 6, para uma tarefa lexicográfica por excelência: a procura de sintagmas nominais cujo núcleo tenha como lema a palavra *"capacidade"*: `/npn[lema,capacidade]`, de forma a averiguar que preposição é usada com esta palavra.

<u CP954-6> Em uma última sala, Júlio Resende mostra uma séria vasta de pequenos formatos, onde **a *capacidade* de interrogar as possibilidades de a pintura** se evidencia, conjuntamente com um humor que pode justificar os consensos que a obra de este pintor geralmente reúne:

<u CP964-4> Para além de o convívio, as jovens atletas puderam praticar o seu desporto favorito e mostrarem **as suas *capacidades*** », afirmou a o PÚBLICO Isabel Cruz, de a Associação de Andebol de Lisboa.
<u CP993-7> De o ponto de vista concorrencial, o agravamento de as perdas de a empresa explica-se, em parte, devido a a « entrada em o mercado de novos estaleiros situados em países de muito baixo custo de mão-de-obra (países de o ex-bloco de Leste), e a aumentos de **a *capacidade* de docagem em áreas bem localizadas relativamente a os grandes fluxos de tráfego marítimo**, além de também disporem de mão-de-obra barata (países de o Médio e Extremo Oriente).
<u CF165-2> Ela foi construída em Marília (cidade onde foi fundada o banco) com ***capacidade* para mil alunos**.
<u CF479-2> A segunda fase de o projeto prevê a construção (para 1995) de o All-Star Music Resort, em o mesmo estilo e com **mesma *capacidade***, que homenageará os estilos musicais de o country, jazz, rock, calypso e musicais de a Broadway.

Figura 6. Extração, com o Águia, de sintagmas nominais com núcleo *capacidade*

O Águia também permite obter uma visão de conjunto sobre a estrutura dos sintagmas que constituem a Floresta, assim como as funções sintáticas que os constituem. Na figura 7 apresentamos o princípio da procura por "Distribuição dos sintagmas constituintes imediatos" de um sintagma nominal:

'art n'	5839
'art n pp'	4743
'art prop'	2925
'n pp'	1832
'art n adj'	1249
'pron-det n'	1115
'n adj'	726
'num n'	616
'art n fcl'	544
'art n adj pp'	495
'art pron-det n'	494
'art adj n'	444
'art adj n pp'	416
'art n pp pp'	390
'art n prop'	351
'art n icl'	298
'pron-det n pp'	287

Figura 7. Como são constituídos os sintagmas nominais na Floresta?

Esperamos ter conseguido, com este pequeno texto, "abrir o apetite" aos leitores para que tentem explorar por si este tesouro para a língua portuguesa, e que possam enriquecê-lo, ajudando-nos a melhorá-lo e a torná-lo mais fácil de utilizar e manipular.

Referências

Afonso, Susana (2004a). “Árvores deitadas: Descrição do formato e das opções de análise na Floresta Sintáctica”. http://www.linguateca.pt/Floresta/ArvoresDeitadas.doc.

Afonso, Susana (2004b). “A Floresta Sintá(c)tica como recurso”. http://www.linguateca.pt/Floresta/Afonso2004Recurso.doc

Afonso, Susana, Eckhard Bick, Renato Haber & Diana Santos (2002a). “Floresta sintá(c) tica: primeiro ano”. http://www.linguateca.pt/Floresta/Afonsoetal2002.rtf.

Afonso, Susana, Eckhard Bick, Renato Haber & Diana Santos (2002b). “"Floresta sintá(c) tica": a treebank for Portuguese”. In M. Rodríguez et al. (eds.), *Proceedings of LREC'2002* (Las Palmas, 29-31 maio 2002), pp. 1698-1703.

Afonso, Susana, Eckhard Bick, Renato Haber & Diana Santos (2002c). “Floresta Sintá(c) tica: um treebank para o português”. In A. Gonçalves & C.N. Correia (eds.), *Actas do XVII Encontro da Associação Portuguesa de Linguística* (APL 2001) (Lisboa, 2-4 outubro 2001). APL, Lisboa, 2002, pp. 533-545.

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

Bick, Eckhard (2003). “A CG & PSG Hybrid Approach to Automatic Corpus Annotation”. In Kiril Simow & Petya Osenova (eds.), *Proceedings of SProLaC2003* (Lancaster, março, 2003), pp. 1-12.

Bick, Eckhard (2004). “Looking at the Floresta Sintá(c)tica with a CorpusEye: A user-friendly cross-language search interface”. http://www.linguateca.pt/Floresta/floresta-corpuseye_en.doc.

Haji?, Jan (1998). “Building a Syntactically Annotated Corpus: The Prague Dependency Treebank”. In *Issues of Valency and Meaning*, Karolinum, Praha 1998, pp. 106-132.

Marchi, Raquel (2004). “Revisão humana da Floresta Sintá(c)tica: exemplos e método”, http://www.linguateca.pt/Floresta/Marchi2004Revisao.doc

Marcus, Mitchell P., Beatrice Santorini & Mary Ann Marcinkiewicz (1993). “Building a large Annotated Corpus of English: The Penn Treebank”. *Computational Linguistics* **19**, Number 2, June 1993, pp. 313-330.

Sampson, Geoffrey (sem data). “SUSANNE Corpus and Analytic Scheme”, last modified June 2004. http://www.grsampson.net/RSue.html [último acesso: 27 de outubro de 2004].

Santos, Diana & Eckhard Bick (2000). “Providing Internet access to Portuguese corpora: the AC/DC project”. In Maria Gavrilidou et al. (ed.), *Proceedings of LREC 2000* (Atenas, 31 maio-2 junho 2000), pp. 205-210.

Santos, Diana & Luís Sarmento (2003). “O projecto AC/DC: acesso a corpora/disponibilização de corpora”. In A. Mendes & T. Freitas (orgs.), *Actas do XVIII Encontro da Associação Portuguesa de Linguística* (Porto, 2-4 outubro 2002), APL, Lisboa, 2003, pp. 705-717.

Santos, Diana (2002). “Um centro de recursos para o processamento computacional do português”. *DataGramaZero – Revista de Ciência da Informação* **3** n.1 fev/02, http://www.dgz.org.br/fev02/Art_02.htm [último acesso: 27 de outubro de 2004]

Santos, Diana (2003). “Timber! Issues in treebank building and use”. In N. J. Mamede, J. Baptista, I. Trancoso & M.G.V. Nunes (eds.), *Computational Processing of the*

Portuguese Language, 6th International Workshop, PROPOR 2003, Faro, 26-27 June 2003, Proceedings, Springer Verlag, 2003, pp. 151-158.

Santos, Diana (no prelo). *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa.*