A Floresta Sintá(c)tica como recurso

Susana Afonso
Projecto Floresta Sintá(c)tica

1. Introdução

O uso de recursos linguísticos parece ser hoje incontestavelmente uma prática cada vez mais necessária em diversos domínios de investigação. O fácil acesso a estes recursos, que, na sua grande parte, se encontram computorizados, constitui um forte motivo para a generalização do seu uso. No entanto, ao contrário da facilidade de acesso e de uso, a construção e disponibilização de recursos linguísticos é uma tarefa muito complexa e não imediata.

Entre os recursos linguísticos, encontram-se dois importantes: corpora e "treebanks". Existem variados tipos de corpora e treebanks para variadas línguas. Para o português, a Linguateca (www.linguateca.pt) constrói e disponibiliza gratuitamente muitos corpora e, numa iniciativa conjunta com o projecto VISL (www.visl.sdu.dk), lança-se na construção do primeiro treebank para o português: Floresta Sintá(c)tica.

Este artigo básico no âmbito da linguística de corpus e da linguística computacional tem como objectivos fornecer uma visão muito geral sobre a construção e importância dos corpora e de treebanks, noções intimamente relacionadas, e insistir na ideia que a construção de recursos linguísticos tão válidos como estes só faz sentido quando eles são utilizados.

O artigo está organizado da seguinte forma: faremos uma pequena introdução à noção de corpus e de "treebank" em geral, focando-nos em particular no processo de construção da Floresta Sintá(c)tica, a partir de dois corpora: CETEMPúblico (Rocha e Santos, 2000) e o CETENFolha. De seguida, a Floresta Sintá(c)tica será descrita como recurso importante, entre outros, para investigação linguística e avaliação de sistemas. Terminaremos esta secção com dois exemplos de como a Floresta Sintá(c)tica pode ser usada naqueles contextos.

2. De corpora a "treebank": processo de construção da Floresta Sintá(c)tica

Um corpus pode geralmente ser definido como um conjunto organizado de itens linguísticos autênticos, isto é, efectivamente produzidos por falantes de uma determinada língua. Um corpus está organizado externamente e internamente. A organização externa diz respeito ao tipo de corpus, ou seja, organização segundo, por exemplo, registo (corpus de texto jornalístico ou literário, ou ainda, corpus falado, dialectal) e tempo (por exemplo, corpus histórico). A organização interna relaciona-se, por outro lado, com decisões sobre a configuração do corpus: não só o tratamento do material linguístico- como, por exemplo, o que incluir/excluir do corpus, atomização (o que se considera uma palavra) e segmentação (critérios de divisão em frases)-, mas também a categorização do material linguístico em extractos por assunto (por exemplo, existência de secções de política, desporto) e toda a parte de numeração/identificação do material.

Com os avanços tecnológicos é hoje possível a compilação e armazenamento electrónicos de corpora de grandes dimensões, obtendo-se corpora organizados de forma sistemática, isto é, à partida, sem inconsistências internas em termos de organização além de facilitar o trabalho do investigador sobre o corpus.

Sendo um corpus uma amostra de linguagem natural, isto é, produzida por falantes reais em contextos reais, é possível o estudo sistemático de um determinado tópico e até a formulação de novas questões para investigação.

Os corpora podem também estar ou não anotados, isto é, analisados linguisticamente por um analisador automático. A análise pode ser semântica, morfológica, sintáctica ou morfossintáctica. A representação do material linguístico também pode ser diversa, dependendo da teoria subjacente à análise automática do corpus.

Os corpora estão na base da construção de "treebanks". Constituirá um "treebank" um determinado corpus que tenha sido analisado automaticamente no formato específico de árvores. Opcionalmente, o "treebank" pode ou não ter sido revisto humanamente. Um exemplo de "treebank" é a Floresta Sintá(c)tica (Afonso et al., 2001; 2002), o primeiro banco de árvores para o português sintacticamente analisadas pelo analisador sintáctico PALAVRAS (Bick, 2000). Os corpora que constituem o "treebank" são o CETEMPúblico (Rocha e Santos, 2000) e o CETENFolha, um dos corpora acessíveis através do projecto AC/DC (Santos & Bick, 2000; Santos & Sarmento, 2002), ambos de texto jornalístico, em português europeu e português do Brasil, respectivamente.

A Floresta Sintá(c)tica tem a particularidade de ter uma parte anotada e revista por linguistas (Bosque), e outra anotada mas não revista (Floresta Virgem). Além disso, o Bosque é disponibilizado integralmente em vários formatos. A presente versão do Bosque (6.1.) contém 7.346 árvores revistas, correspondendo a 1468 extractos, 7.283 frases distintas, 174.475 unidades, aprox. 150.057 palavras

O processo de construção deste "treebank" para o português é o seguinte: os corpora são anotados em primeiro lugar em formato de "Gramática Restritiva" (CG, "Constraint Grammar") e o resultado da análise automática revista. A anotação em formato CG inclui lemas, etiquetas de morfologia, categoria gramatical e sintaxe, e etiquetas que embora secundárias são importantes em termos de especificação mais fina das categorias (como veremos na secção sobre a avaliação de sistemas). O seguinte exemplo ilustra o resultado da análise e revisão do formato CG:

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

O exemplo acima é um caso em que não só o conhecimento linguístico mas também extra-linguístico intervêm no processo de revisão. A revisão de *Acontece* e de *Zombi dos Palmares* decorre do conhecimento extra-linguístico: o primeiro, um programa da TV2 portuguesa, e o segundo uma personagem histórica brasileira. Daí a análise automática ter de ser alterada nos seguintes níveis: lema, categoria gramatical,

morfologia, sintaxe, e no segundo caso atomização. Relativamente à atomização, é importante referir que, na Floresta Sintá(c)tica, os nomes próprios (sejam eles, topónimos, antropónimos, entre outros) são considerados uma única unidade, daí a análise de *Zombi dos Palmares* como uma unidade e não decomposta nas suas partes. Consequentemente, o segmento que segue *Acontece* passa a ter uma análise sintáctica de dependente nominal e não de adverbial, uma vez que a análise verbal de *Acontece* foi também ela alterada.

É a partir da revisão de CG que o formato de árvores é criado. Este formato contém a mesma informação que o formato CG, incluindo adicionalmente os níveis de constituintes, indicados por sinais de igual:

```
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
```

As árvores são também revistas em todos os seus níveis de análise. Só após essa revisão, as árvores passam a integrar o Bosque.

O processo de revisão é bastante complexo, na medida em que todos os níveis de análise são revistos: lema, morfologia, categoria gramatical, sintaxe, etiquetas secundárias e, no caso do formato de árvores, nível de constituintes, forma dos nós não-terminais (isto é, nós com dependentes) e função e forma da raiz. O processo de revisão do formato de árvores tem, por vezes, efeitos retroactivos: a detecção de problemas na análise das árvores conduz a uma revisão parcial do formato CG, uma vez que o formato de árvores provém da revisão do formato CG.

O formato de árvores do Bosque tem associada uma extensa documentação (Afonso, 2004), em contínuo desenvolvimento, resultado das discussões conjuntas com os restantes membros do projecto durante a construção da Floresta Sintá(c)tica e que inclui a descrição das etiquetas utilizadas e o próprio formato de árvores, mas também as escolhas que foram sendo feitas ao longo do projecto.

As árvores do Bosque podem ser interrogadas através de dois sistemas de busca publicamente acessíveis online (Águia e tgrepeye), que permitem a extracção de ocorrências de um determinado fenómeno linguístico pedido bem como o cálculo da sua frequência. O pedido de busca pode ser feito por itens lexicais ou com base nas etiquetas de anotação. Neste último caso, a busca nas árvores do Bosque torna-se particularmente relevante, uma vez que, sendo as análises do Bosque revistas por linguistas, a extracção de padrões não desejados, será à partida menor.

A Floresta Sintá(c)tica é um recurso criado a fim de ser usado por diferentes utilizadores para diversos fins, entre os quais destacamos quatro: ensino, descrição linguística, treino de analisadores morfossintácticos e avaliação de sistemas. Vejamos alguns exemplos concretos de uso da Floresta Sintá(c)tica em termos de descrição linguística e avaliação de sistemas.

3. Floresta Sintá(c)tica como recurso para descrição linguística

Suponhamos que se queria investigar a estrutura argumental dos verbos em português. Os seguintes exemplos ilustram o tipo de objecto em estudo:

Tipos de verbos	Exemplos
Verbos intransitivos	CP14-3 Um endurecimento nítido **existe** desde então neste terreno altamente perigoso.
Verbos transitivos	CF2-5 Manchete **estréia** novo jornalístico
Verbos ditransitivos	CF14-2 Souza também **negou** aos réus o direito de apelarem da setença em liberdade. CF57-3 Com medo de perder poderes, Modiano chamou Mourão e, numa atitude mesquinha, **comunicou**-lhe que, como o financiamento do projeto seria bancado por o BNDES, ele iria para lá, mas na qualidade de representante do banco.

A nossa investigação basear-se-á fundamentalmente em dois aspectos: 1) quantas frases no Bosque exibem verbos intransitivos, transitivos e ditransitivos, e 2) relativamente aos argumentos verbais, quais as formas possíveis (argumentos verbias que são orações, sintagmas, pronomes, etc.) e qual a frequência no Bosque das formas que cada argumento exibe.

Como os verbos na Floresta Sintá(c)tica não estão marcados quanto à sua transitividade, as extracções a partir do Águia terão de ser realizadas a partir da informação sintáctica dos argumentos, no caso dos verbos transitivos e ditransitivos, já que os verbos intransitivos não possuem argumentos. Assim, de forma a se extrair frases com verbos transitivos, por exemplo, é necessária a presença do objecto directo (ACC) na expressão de busca, por exemplo: ass_fcl('.*P ACC.*) (oração finita cuja estrutura interna contenha pelo menos um verbo principal (P) e um objecto directo (ACC), adjacentes ou não). O mesmo tipo de busca poderia ser realizado para orações não finitas, de forma a extrair ocorrências como: *CF4-1 Em a volta de uma viagem ao exterior, vale a pena* ***trazer*** *uma impressora matricial*. Para extrair frases contendo verbos ditransitivos, o mesmo tipo de expressão de busca seria usado, incluindo mais um tipo de argumento: objecto indirecto pronominal (DAT), como em CF57-3 no quadro acima, ou preposicional (PIV), como em CF14-2, no quadro acima.

De forma a investigar os tipos de forma possíveis que os argumentos exibem, é possível obter apenas as frequências, bastando realizar a procura por distribuição de forma. É possível também obter as concordâncias correspondentes para inspecção.

Relativamente aos verbos intransitivos, que não possuem estrutura argumental, o tipo de busca conta com outro tipo de elementos, como procura de verbos finitos que ocorram isoladamente, ou verbos finitos ou não-finitos que co-ocorram com adjuntos adverbais (ADVL), que por definição não são argumentos verbais.

Expomos aqui alguns resultados das buscas:

Tipos de frases, segundo transitividade	Frequências no Bosque	Exemplos de concordância

Frases com verbos intransitivos[1] finitos Frases com verbos intransitivos[2] não-finitos	701 638	CF151-32 É mole mas **sobe**! CF272-1 **Cresce** o volume de negócios em a Bovespa CP119-3 A GF confiscou ainda uma viatura ligeira de marca Bedford, envolvida em a rede de contrabando, ontem **descoberta**.
Frases com verbos transitivos (verbo e argumento adjacentes) • Verbo finito • Verbo não-finito	 5002 1634	 CF467-15 Eu me **tirei** de as ruas e comecei a estudar. CF482-2 A Folha realizou um teste em um de os ônibus que integram a nova linha Penha-lapa, **acompanhando** a décima viagem de ontem.
Frases com verbos ditransitivos • Verbo finito e objecto indirecto pronominal adjacentes • Verbo não-finito e objecto indirecto pronominal adjacentes • Verbo finito seguido de objecto directo e objecto indirecto pronominal adjacentes	 154 28 1	 CP52-3 Sempre me **pareceu** estranho nunca ter lido um artigo sobre a « ilha de Santos ». CP689-2 O presidente desculpou-se, **lembrando**-lhe, mais uma vez, que tudo aquilo é provisório. CP243-11 Quando o navio estava a passar em o estreito de os Dardanelos, ..., **apresenta**-se-lhe uma velha senhora magra, elegante e «pripudrennaia» ...

As formas que, por exemplo, o objecto directo pode exibir encontram-se na seguinte tabela:

funções	formas	Frequências
ACC	Sintagma nominal (np)	4683
	Oração finita (fcl)	1172
	Oração não-finita (icl)	545
	Sintagma adjectival (ap)	19
	Sintagma preposicional (pp)	9
	Oração deverbal (acl)	8
	Sintagma adverbial (advp)	2

[1] As estruturas procuradas foram, com ordens dos constuituintes diversas: a)Predicador (P); b) Sujeito (SUBJ) Predicador; c) P Adjunto Adverbial (ADVL); d) SUBJ P ADVL.
[2] As estruturas procuradas foram, com ordens dos constuituintes diversas: a)Predicador (P); b) Sujeito (SUBJ) Predicador; c) P Adjunto Adverbial (ADVL); d) SUBJ P ADVL.

É importante, no entanto, realçar que o cálculo das frequências obtidas automaticamente, é aproximada, isto é, para se obter valores reais é necessário inspeccionar manualmente as concordâncias extraídas.

No estudo acima, a inspecção manual torna-se fundamental no caso do objectos indirectos preposicionais. Na Floresta Sintá(c)tica, os objectos indirectos preposicionais não possuem uma etiqueta exclusiva da sua função, mas partilha a etiqueta com outros objectos que tenham forma preposicional (PIV). Daí que a busca baseada na etiqueta PIV irá extrair todos os objectos preposicionais, incluindo o objecto indirecto, sendo fundamental a inspecção manual e selecção dos casos relevantes. Por esta razão, não apresentamos os valores na tabela relativos ao objecto indirecto preposicional.

Muitos outros aspectos relacionados com a estrutura argumental do verbos poderiam ser ainda explorados, a partir da inspecção das concordâncias extraídas, como, por exemplo, estabelecer contextos sintáctico-semânticos em que determinada ordem do verbo e seu(s) argumento(s) pode ocorrer, à semelhança do estudo realizado para a coordenação de avérbios em –mente (Afonso, 2002).

A vantagem em usar a Floresta Sintá(c)tica é o facto de se poderem realizar buscas sobre funções sintácticas e formas e não apenas sobre itens lexicais, o que impossibilitaria o estudo que aqui propomos. As buscas no Bosque, em particular, são qualitativamente superiores, porque o Bosque corresponde à parte do treebank revista por linguistas. Em termos de ensino, um tipo de corpus como o Bosque é ideal. Por outro lado, estudos exclusivamente quantitativos a partir do Bosque deverão ter conta o seu actual tamanho reduzido na fase das conclusões do estudo.

4. Floresta Sintá(c)tica como recurso para avaliação de sistemas

Além da investigação linguística, a Floresta Sitá(c)tica, e especialmente o Bosque, pode ser um ferramenta muito útil para avaliar o desempenho de sistemas em anotação automática de corpora.

Tomemos como exemplo a primeira iniciativa de avaliação conjunta para o português (Santos, no prelo). De forma a avaliar o desempenho em anotação morfológica dos sistemas participantes nesta iniciativa foi necessário criar uma lista dourada, ou seja, “uma lista de formas recolhida de forma organizada que foram classificadas e revistas por diferentes anotadores humanos e seguindo as directivas da comissão organizadora e científica” (Barreiro e Afonso, no prelo). Foi a partir desta lista que os sistemas foram avaliados em termos de análise morfológica.

Uma espécie de “lista dourada”, apesar de classificada automaticamente, mas revista por diferentes anotadores humanos, é o Bosque. O Bosque poderia constituir uma base para a avaliação de sistemas em diferentes níveis pelos seguintes motivos: além da revisto por linguistas, o Bosque 1) é rico em informação, podendo ser usado em avaliações de morfologia e categoria gramatical, lema, sintaxe e níveis de constituintes, 2) é flexível, tanto em termos de formatos em que pode ser disponibilizado como na possibilidade de simplificação de etiquetas.

A flexibilidade de formatos e etiquetas é uma questão importante para avaliação, na medida em que o padrão base de avaliação terá de possuir um conjunto de etiquetas e formato compatíveis com os dos sistemas a avaliar, tal como aconteceu com a lista dourada usada para as Morfolimpíadas.

Um exemplo de simplificação de etiquetas já efectuado no Bosque, foi a transformação de *como* e *segundo*, em preposições, nos seguintes contextos:

CP44-2 ***Segundo* fontes policiais** citadas pelo «New York Times», o atentado poderia ter partido da velha guarda do clã Gambino.
CF65-6 ***Como* notou o jornalista Jim Greer**, da revista americana «Spin», é uma boa grana, mas nada muito absurdo.

A simplificação de etiquetas é possibilitada pela existência de etiquetas secundárias que adicionam informação sobre as funções específicas dos constituintes em determinados contextos. Nos exemplos *segundo* e *como*, o procedimento normal do anotador automático era manter a categoria gramatical de advérbio e adicionar a informação de preposição como etiqueta secundária, indicando que, naqueles contextos, o constituinte funcionava como uma preposição. Assim, simplificou-se o sistema de anotação, eliminando a informação secundária ao torná-la "principal", ou seja, passando a ser a categoria gramatical da palavra.

5. Conclusão

A Floresta Sintá(c)tica é um recurso usável, como esperamos ter demonstrado, através de exemplos concretos de uso da Floresta para descrição linguística e avaliação de sistemas. Outras áreas que podem beneficiar do uso deste recurso poderiam aqui ter sido exploradas, entre as quais mencionamos o treino de analisadores automáticos e ensino. Relativamente a estes dois pontos, refira-se que a Floresta Sintá(c)tica não é um treebank rígido, mas flexível, possibilitando a filtragem de etiquetas, além da da sua disponibilização em diversos formatos. Quanto ao ensino, o recurso a corpora e treebanks para construção de materiais que possam ser usados em contexto de sala de aula é uma realidade no projecto VISL, cujo sistema é o adoptado nas escolas públicas na Dinamarca.

A Floresta Sintá(c)tica, em particular, é um projecto em constante crescimento tanto em quantidade (à medida que mais árvores vão sendo revistas) como em qualidade (à medida que melhores análises vão sendo implementadas e mais ampla documentação vai sendo produzida). O uso do recurso teria apenas a beneficiar todos os envolvidos no processo, construtores e utilizadores; o resultado do feedback sobre os eventuais problemas de utilização seria o aperfeiçoamento continuado da Floresta.

Bibliografia:

Afonso, S. (2002). "Clara e sucintamente: um estudo em corpus sobre a coordenação de advérbios em -mente". In Amália Mendes & Tiago Freitas (orgs.), *Actas do XVIII Encontro da Associação Portuguesa de Linguística (APL 2002)* (Porto, 2-4 Outubro 2002), Lisboa: APL, pp. 27-36.

Afonso, S. (2004). Árvores deitadas: Descrição do formato e descrição das opções de análise na Floresta Sintáctica. Texto produzido no âmbito da Floresta Sintá(c)tica.

Afonso, S., Eckhard Bick, Renato Haber & Diana Santos (2001). "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, 2002. pp. 533-545.

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

Barreiro, A. & Susana Afonso (no prelo). "Construção da lista dourada para as primeiras Morfolimpíadas do português". In Diana Santos (org.). *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*.

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

Rocha, P. & Diana Santos (2000). "CETEMPúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa". In Maria das Graças Volpe Nunes (ed.), V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR 2000) (Atibaia, SP, 19 a 22 novembro de 2000), São Paulo: ICMC/USP, pp. 131-140.

Santos, D. & Eckhard Bick (2000). "Providing Internet access to Portuguese corpora: the AC/DC project". In Maria Gavrilidou et al. (ed.), *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC 2000)* (Athens, 31 May-2 June 2000), pp. 205-210

Santos, D. & Luís Sarmento (2002). "O projecto AC/DC: acesso a corpora/disponibilização de corpora". In Amália Mendes & Tiago Freitas (orgs.), *Actas do XVIII Encontro da Associação Portuguesa de Linguística (APL 2002)* (Porto, 2-4 Outubro 2002), Lisboa: APL, pp. 705-717.

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