Floresta Sintá(c)tica: Ficção ou realidade?

Eckhard Bick, Diana Santos, Susana Afonso e Raquel Marchi

Neste artigo descrevemos o primeiro conjunto de árvores sintácticas revistas para o português, a Floresta Sintá(c)tica, dando especial ênfase à sua possível utilização num modelo de avaliação conjunta.

Após uma descrição dos vários usos de um conjunto de árvores como este (designado em inglês por "treebank"), na secção 1, e uma apresentação do processo de construção do recurso, na secção 2, a secção 3 é dedicada à apresentação da Floresta para avaliação conjunta, após uma breve inspecção da literatura sobre os usos deste tipo de recursos nesse âmbito. Terminamos o artigo, na secção 4, com uma chamada às armas no sentido de fomentar uma maior participação na florestação da língua portuguesa.

1. Usos e razões de uma Floresta

Uma floresta sintáctica pode ser considerada do ponto de vista da linguística descritiva ou do ponto de vista da engenharia da linguagem, donde, os seus principais usos diferem conforme cada um destes pontos de vista, como ilustrado na Figura 24-1: enquanto a linguística descritiva vê uma floresta como favorecendo sobretudo o ensino e a extracção de dados quantitativos, ou estatísticas, sobre a realidade da língua, um linguista computacional utiliza uma floresta como uma ferramenta para treinar e medir o seu analisador sintáctico. Na secção 3, apresentaremos ainda outra forma de olhar para a nossa Floresta, nomeadamente no âmbito da organização de avaliações conjuntas.

Por vezes, a razão para a criação de "treebanks" pode ter sido o desejo de instanciar uma dada teoria linguística, ou melhor, os seus aspectos estruturais, ao nível sintáctico e/ou semântico. Embora a maior parte das florestas de tamanho razoável se auto-considerem corpora de referência para análise sintáctica generalizada, para uma dada língua, é difícil agradar a todos os utilizadores, e é impossível evitar totalmente a orientação ou influência de uma teoria linguística, por mais neutro que se deseje ser.

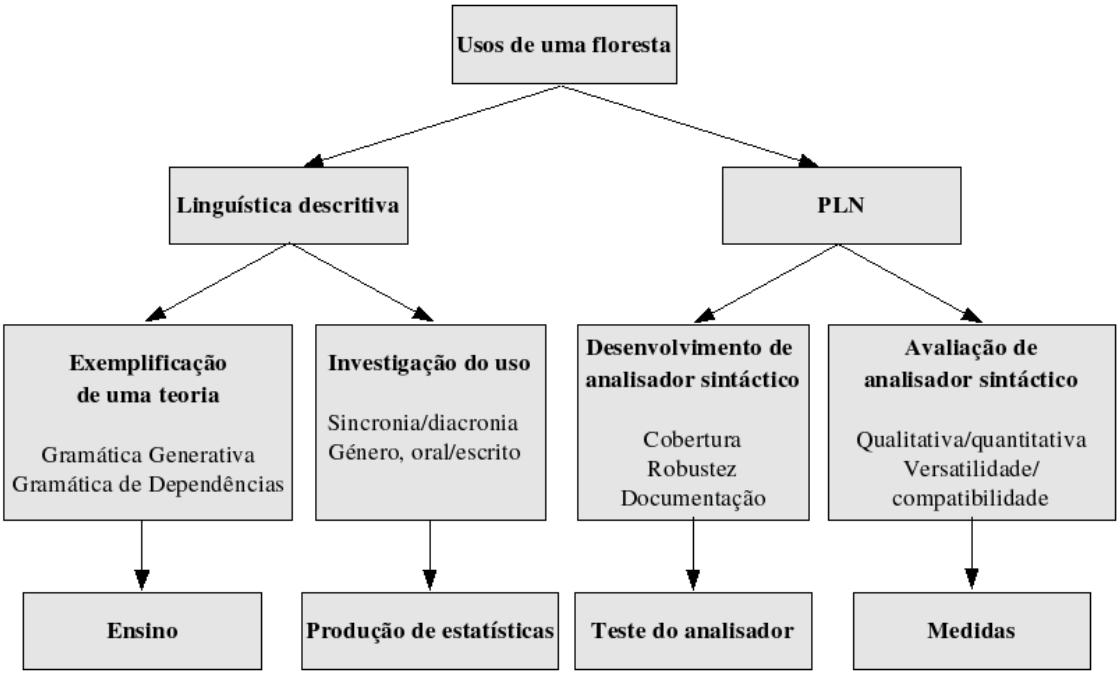

Figura **1-1**: Várias perspectivas sobre uma floresta

As florestas "clássicas" como o Penn Treebank (Marcus *et al.*, 1993, Taylor *et al.*, 2003) e SUSANNE (Sampson, 1995), são conceptualmente baseadas na descrição de estrutura através de agrupamento (por chavetas, "bracketing"), enriquecida com anotação funcional,

enquanto que a gramática dependencial subjaz a várias outras iniciativas mais recentes, tal como o grande Prague Dependency Treebank (PDT) para o checo (Hajič, 1998), e outras florestas mais pequenas para o turco, o russo, o dinamarquês e o italiano.

Dado que uma floresta grande beneficiaria obviamente de uma análise sintáctica automática, não é de espantar que ferramentas de PLN interajam intimamente com questões de descrição da língua. Assim, nas florestas Alpino (para o holandês) e BulTreeBank (para o búlgaro) foi usado o formalismo gramatical Head-driven Phrase Structure Grammar (HPSG, Pollard & Sag, 1994), enquanto que na floresta UAM, para o castelhano, foi usada a LFG (Kaplan & Bresnan, 1982). Um dos módulos do TIGER (Brants *et al.*, 2002), para o alemão, também se baseia neste formalismo. A floresta TIGER, de facto, é um projecto híbrido (sintagmático e dependencial).

Também a Floresta Sintá(c)tica vem em dois formalismos paralelos: (a) uma floresta dependencial com anotação de gramática de restrições (CG) por palavra, e (b) uma floresta sintagmática com anotação de constituintes. Ambas as versões incluem a especificação de descontinuidades (ou ramos cruzados), e ambas especificam função sintáctica e forma sintáctica. Podemos pensar nas duas versões como reflectindo dois tipos de análise sintáctica: anotação ("tagging") e agrupamento ("chunking"). A primeira descreve a estrutura sintáctica como uma relação entre palavras (traços dependenciais), a segunda caracteriza essa mesma estrutura como uma hierarquia de constituintes (ou seja, grupos de palavras, sintagmas). Construímos uma floresta a partir da outra (agrupando depois de anotar), preservando a função sintáctica. Desta forma, esperamos conseguir o melhor dos dois mundos, optimizando o "potencial de filtragem", ou seja uma compatibilidade máxima com diferentes teorias e notações sintácticas. Além disso, do ponto de vista prático, ao fazer uma análise (e consequente revisão) em dois estágios, não só aumentamos a qualidade do recurso, como aproveitamos ao máximo a conhecida robustez da gramática de restrições, enriquecendo-a com a representação em níveis de constituintes.

A nossa solução para o problema da compatibilidade teórica das florestas é, pois, a criação de transdutores, ou filtros, para cada variedade individual de formalismos pertencentes às duas grandes famílias acima mencionadas. Por exemplo, mencione-se o programa criado pelo primeiro autor para transformar a Floresta numa "irmã" do Penn Treebank.

Evidentemente, nem todas as diferenças em linguística teórica podem ser obtidas através de uma manipulação (ou filtragem) do que está presente na Floresta Sintá(c)tica. Mas, em geral, pode dizer-se que é sempre mais fácil remover (ou aglomerar) categorias demasiado finas do que adicionar novas categorias ou distinções. Por isso, e para aumentar a generalidade do recurso criado e satisfazer os seus utilizadores, a equipa da Floresta Sintá(c)tica tem várias vezes adicionado novas distinções mais finas para possibilitar futuras filtragens. Um exemplo é a distinção entre AVDL (adjunto adverbial livre), ADVS e ADVO (complementos com valor adverbial, referente ao sujeito e ao objecto directo, respectivamente). A distinção baseia-se no tipo de comportamento sintáctico que ADVL, por um lado, e ADVS e ADVO, por outro, exibem. Assim, ao contrário dos complementos marcados ADVL, os marcados como ADVS ou ADVO exibem um comportamento sintáctico mais próximo dos argumentos do que dos adjuntos adverbiais livres, por serem de presença obrigatória. No entanto, semanticamente assemelham-se a adverbiais, expressando tempo, lugar, modo, etc. Os seguintes exemplos ilustram o uso das funções ADVL, ADVS e ADVO, respectivamente:

CP23-8 Há, **no ar**, uma certa ideia de invasão.

CP18-8 Com o dilúvio como pano de fundo, o empate traduz de forma feliz um jogo que ficou **no meio**.

CP66-7 Daniel Blaufuks apresenta um trabalho construído a partir de imagens previamente existentes, o que situa este trabalho **na linha das mais recentes preocupações do autor (a questão da perda de referência das imagens fotográficas, o trabalho sobre a simulação)**, não deixando contudo de estar sempre presente o registo poético, em parte autobiográfico, que sempre o caracterizou:

Uma aplicação directa e óbvia do desenvolvimento da Floresta Sintá(c)tica é a melhoria de analisadores sintácticos com base no material, e a melhoria do próprio analisador sintáctico original, ou inicial, o PALAVRAS (Bick, 2000). No entanto, dado o problema dos dados

esparsos, e o facto de a Floresta não ter sido primordialmente desenvolvida para essa função (apenas), várias questões têm de ser resolvidas antes de empregar a Floresta com esse objectivo. Por exemplo, seria aconselhável a criação de um filtro que amalgamasse forma e função numa única etiqueta, assim como um mapeamento em etiquetas menos finas e numerosas, como é usado no VISL-lite (Bick, 2003) com objectivos pedagógicos.

Por outro lado, é importante referir que, para efeitos estatísticos e mais uma vez para atingir um número maior de beneficiários, a Floresta pode ser considerada como um objecto maior, incluindo também a Floresta Virgem, que corresponde a um milhão de palavras (em cada uma das duas variantes do português) analisadas automaticamente pela última versão do analisador usado[1], mas ainda não revistas e depuradas.

2. O recurso "Floresta Sintá(c)tica"

O projecto da Floresta Sintá(c)tica tem como objectivo principal a construção de um "treebank" enquanto recurso linguístico em língua portuguesa, publicamente distribuído, que possa ser usado por diferentes utilizadores para diferentes fins, tais como o ensino do português e da sintaxe portuguesa, a descrição linguística, o treino de analisadores morfossintácticos e a avaliação de sistemas.

O recurso produzido no projecto Floresta Sintá(c)tica inclui o que chamamos a Floresta Virgem (ainda não desbravada e revista pelos linguistas) e o Bosque, a aumentar continuamente, e que corresponde à parte revista e verificada da Floresta.[2]

Para que possa chegar a um número elevado de potenciais utilizadores, tem havido um esforço de documentação apreciável associado a este recurso (no que se refere ao Bosque), como o comprovam os textos Afonso *et al.* (2002a,b,c), de apresentação do projecto, Santos *et al.* (2004), que põe a ênfase na parte pedagógica e Vilela *et al.* (2005) na divulgação para um público informático. Destaque-se também uma série de apresentações, disponíveis do sítio da Floresta, para disseminar o recurso e dar exemplos variados da sua utilização. Além disso, o formato de árvores do Bosque tem ainda associada uma extensa documentação em contínuo desenvolvimento (Afonso, 2004-2006), 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.

Na Tabela 1-1, caracterizamos o Bosque na sua presente versão, a versão 6.4.

Tabela 1-1: Descrição de algumas estatísticas referentes ao Bosque

Palavras	~175.000
Unidades	~200.000
Frases	8.755
Árvores	8,818
Orações	21.042
Orações finitas	14.637
Orações infinitivas	5.359
Orações averbais	1.406
Sintagmas nominais	40.671
Sintagmas preposicionais	30.363
Sintagmas adjectivais	1.720
Sintagmas adverbiais	790
Itens coordenados	5.214
Sintagmas descontínuos	282

[1] Ao contrário de outros projectos, em que, uma vez todo o material automaticamente analisado, se passa a 100% para a análise intelectual (humana), o desenvolvimento da Floresta, ao levar à imediata melhoria do PALAVRAS, faz com que o material a rever seja sucessivamente analisado por novas versões do analisador automático, que levam imediatamente em conta a correcção de possíveis erros e a introdução de novas distinções motivadas pelo desenvolvimento da Floresta.

[2] Existe alguma instabilidade na definição de *treebank* em inglês e consequentemente *floresta* em português: embora uma verdadeira floresta tenha de ser revista e verificada, muitas vezes usa-se essa designação para referir somente um conjunto de árvores sintácticas.

O material que compõe a Floresta vem de dois corpora maiores, em português de Portugal (CETEMPúblico) e em português do Brasil (CETENFolha), ambos de texto jornalístico, anotados automaticamente pelo analisador morfossintáctico PALAVRAS.

2.1 Historial e resumo da metodologia

A Tabela 1-2 esboça o percurso seguido no projecto Floresta Sintáctica até ao presente.

Tabela 1-2: Calendário do projecto Floresta Sintá(c)tica

Novembro de 2000	Início do projecto
Janeiro 2001	Apelo à comunidade para participar
Até Abril 2001	Preparação do material textual
Setembro de 2001	Anúncio da Floresta na lista corpora
Dezembro de 2001	Primeira fase concluída, 2000 árvores
Dezembro de 2002	Adição de português brasileiro
Março 2004	Primeira versão com floresta brasileira
Junho 2004	Floresta distribuída no formato actual

Desde Dezembro de 2001, a Floresta tem continuado a progredir lentamente, com os elementos da equipa (os autores do artigo) trabalhando em quatro localizações diferentes. Muito sucintamente, a forma como o trabalho se desenrola presentemente é a revisão de um conjunto razoável de frases (novas, ou revisão da versão anterior do Bosque) e sua instalação no pólo da Linguateca de Oslo, que faz uma primeira validação e actualiza o sítio onde se acede à Floresta. Essa revisão pode levar ao levantamento de dúvidas (discutidas pela equipa) e à alteração da documentação, assim como a uma posterior nova revisão, e é seguida pela instalação da Floresta no projeto VISL. A partir de Agosto de 2004, o pólo de Braga da Linguateca fornece uma validação adicional, e cria ainda outras versões da Floresta, em formatos mais internacionais, para mais fácil disseminação, acessíveis de outro ponto da rede.

Do ponto de vista do trabalho linguístico, contudo, há apenas dois formatos relevantes: o da CG e o das árvores deitadas; os outros são variações sistemáticas (e automáticas) do último. A Figura 1-2 apresenta duas frases nos dois formatos.

<ext n=1 sec=clt sem=92b> <s> O [o] <artd> DET M S @>N 7=e=Meio [7=e=Meio] PROP M S @SUBJ> é [ser] <fmc> V PR 3S IND VFIN @FMV um [um] <arti> DET M S @>N ex-libris [ex-libris] N M P @<SC de [de] <sam-> PRP @N< a [o] <-sam> <artd> DET F S @>N noite [noite] N F S @P< algarvia [algarvio] ADJ F S @N< $. </s>	SOURCE: CETEMPúblico n=1 sec=clt sem=92b CP1-2 O 7 e Meio é um ex-libris da noite algarvia. A1 STA:fcl =SUBJ:np ==>N:art('o' <artd> M S) O ==H:prop('7_e_Meio' M S) 7_e_Meio =P:v-fin('ser' PR 3S IND) é =SC:np ==>N:art('um' <arti> M S) um ==H:n('ex-libris' M P) ex-libris ==N<:pp ===H:prp('de' <sam->) de ===P<:np ====>N:art('o' <-sam> <artd> S) a ====H:n('noite' F S) noite ====N<:adj('algarvio' F S) algarvia .
<ext id=17 cad="Ilustrada" sec="nd" sem="94b"> <s> O [o] <artd> DET M S @>N panorama [panorama] N M S @SUBJ> sofre [sofrer] <fmc> V PR 3S IND VFIN @FMV prejuízos [prejuízo] N M P @<ACC demais [demais] <quant> DET M P @N<	SOURCE: CETENFolha n=17 cad="Ilustrada" sec="nd" sem="94b" CF17-5 O panorama sofre prejuízos demais em favor da tese. A1 STA:fcl =SUBJ:np ==>N:art('o' <artd> M S) O ==H:n('panorama' M S) panorama =P:v-fin('sofrer' PR 3S IND) sofre =ACC:np

em=favor=de [em=favor=de] <sam-> PRP @<ADVL a [o] <artd> <-sam> DET F S @>N tese [tese] N F S @P< $. </s>	==H:n('prejuízo' M P) prejuízos ==N<:pron-det('demais' <quant> M P) demais =ADVL:pp ==H:prp('em_favor_de' <sam->) em_favor_de ==P<:np ===>N:art('o' <artd> <-sam> F S) a ===H:n('tese' F S) tese .

Figura 1-2: Duas frases, no formato CG e no formato AD

2.2 Problemas de interpretação

Como seria de esperar, dado que a anotação é bastante complexa, também o é a sua revisão. Assim, existem muitos casos em que não há uma resposta unânime para a resolução de um dado problema. Em vez de tentar argumentar *ad eternum*, a nossa abordagem tem sido a de escolher a alternativa que, por um lado, envolva menos trabalho para o anotador e, por outro, seja mais fácil de recuperar (caso haja uma mudança de opinião). A principal resposta, contudo, tem sido a de tentar documentar claramente as opções tomadas, sobretudo se já foram controversas no seio da equipa. Basta folhear a nossa "bíblia florestal" (Afonso, 2004a) para ficar com uma ideia da miríade de pequenas decisões que é preciso tomar quando o objectivo é analisar todo o texto e não apenas fenómenos da nossa eleição.[3]

Por outro lado, casos há em que uma distinção razoavelmente clara se esbate em muitos contextos, como é o caso já amplamente discutido (Santos, 2003) da diferença entre as funções adjecto predicativo (N<PRED) e aposto (APP), ambas caracterizando dependentes nominais,[4] respectivamente ilustradas pelas frases seguintes:

> CP3-2 O primeiro fabricante mundial de «ratos» para computador, **a empresa suíça Logitech**,…
> CF11-3 O Applause, **um sedã quatro portas, com motor 1.6**, é o carro mais caro da Daihatsu.

Se não se aboliu esta distinção na Floresta foi por se considerar que seria deitar fora o trabalho do analisador sintáctico que, nos casos mais flagrantes, acertava, enquanto que o seu desempenho nos casos cinzentos, dado que ambas as alternativas eram aceitáveis, não se podia considerar errado.

De qualquer forma, e como já argumentado em Santos (1996) a propósito da construção de um analisador morfológico, existem sempre fenómenos de fronteira. Outro exemplo: dado que os substantivos comuns e os próprios são classificados com uma categoria gramatical diferente (N ou PROP), qual a melhor escolha para *Governo* em CP29-4 ou CP919-2?

> CP29-4 O **Governo** gabonês, num comunicado em que dá conta da existência de «uma epidemia» na região, pede aos habitantes que não evacuem os doentes nem para a capital da província, Makokou, nem para a capital nacional, Libreville, e que alertem para qualquer novo caso as autoridades sanitárias, para que estas possam providenciar o tratamento dos doentes «in loco».
> CP886-3 A aplicação aos aviões comerciais de passageiros dos sistemas electrónicos de defesa antimíssil, em uso na aviação militar, está a ser equacionada pelo **governo** norte-americano.
> CP919-2 R. – Até agora, a política do **Governo** para as privatizações tem sido orientada de modo a obter o melhor encaixe financeiro.

Outra situação que é preciso mencionar é precisamente a oposta à discordância, mas que também é frequente: a inexistência de qualquer opinião sobre como analisar um dado fenómeno, proveniente tanto do facto de essas questões serem sempre omissas em gramáticas como pela incerteza em relação a que alternativa pode acautelar melhor os interesses de um utilizador genérico.

---

[3] Para os leitores que suspeitem que esta constatação reflicta sobretudo falta de conhecimento de linguística e gramática por parte da equipa, recomenda-se a leitura de Sampson (2003) sobre as prioridades de uma floresta em contraposição à sintaxe teórica.

[4] O aposto pressupõe uma relação de identidade relativamente ao núcleo, enquanto o adjecto predicativo pressupõe uma relação de predicação, adicionando informação sobre o núcleo que modifica.

Citamos o caso das percentagens, das expressões do tipo *de X para Y* e de usos de palavras tradicionalmente consideradas advérbios como preposições (*segundo*, *como*) ou conjunções coordenativas (*além de*, *mais*). Exemplos:

CP40-5 O fim da guerra fria, com a implosão da URSS acabou com a importância helvética no tabuleiro europeu, enquanto os recentes escândalos de cumplicidade com os nazis, **mais** o roubo das economias dos judeus pelos seus bancos lhe tiraram a simpatia americana.

CP93-2 Cerca de um milhão de contos em diamantes, em gemas e para uso industrial, foi roubado na noite **de 14 para 15** de Setembro da estação de escolha da Sociedade Portuguesa de Empreendimentos (SPE), no Ocapa, Lunda Norte.

CP149-6 Os dois criminosos tinham libertado, ontem, por volta das 4h50 locais, os três reféns que detinham, um dos quais ferido, **segundo** a polícia de Wiesbaden.

CF60-2 Em outro espaço, há exibição de vídeos que demonstram, por exemplo, a cerimônia de o chá e o teatro kabuki, **além de** pontos turísticos do Japão.

Outros casos, mais complicados, são verbos com valor de advérbio, que representam um desafio para a representação sintáctica, uma vez que constituem marcadores discursivos, e as construções parentéticas, cujas últimas abonações demonstram claramente que o texto jornalístico não inclui apenas informação a seco, mas exibe criatividade e muito mais:

CF758-4 «Eu torci muito por você, **viu**?», disse Itamar a Fu.

CF821-1 Niemeyer – **É**, foi em 1936.

CF372-1 Enquanto isso **(ou entrementes, como nas antigas histórias em quadrinho)**, o ex-governador ACM descola suntuoso empréstimo para socorrer os cacaueiros da Bahia, quatro anos de carência, juros de 2% subsidiados, um empréstimo de pai (governo) para filho (fazendeiros).

CF372-3 Enquanto isso **(outra vez a vontade de escrever entrementes)**, a Receita Federal divulga a lista dos devedores de o Imposto de Renda que nada pagarão.

Finalmente, uma das decisões que tomámos desde o início foi a de usar a nossa compreensão como falantes do português para podar árvores e interpretações sintácticas espúrias. Contudo, é interessante verificar que existem ainda 121 frases com mais de uma interpretação no material do Bosque, ou seja, casos de ambiguidade ou vagueza que foram explicitamente reconhecidos pelo anotador.

De facto, é possível especificar árvores alternativas para uma mesma frase, ou simplesmente codificar, numa mesma árvore, alternativas de ligação ("attachment"). Tal pode corresponder a duas situações: a) ambiguidade (duas ou mais interpretações), normalmente acompanhada por representações sintácticas distintas, e b) duas ou mais configurações sintácticas para um mesmo fenómeno, sobre o qual não existe ainda consenso para preferir uma das representações dadas como possíveis. Como exemplo deste último caso, veja-se:

CP46-1 Na mesma ocasião iniciaram-se investigações que incidiram sobre o árbitro madeirense Marques da Silva, **também ele suspeito de se ter deixado corromper**.

Em CP46-1, existem duas configurações sintácticas para o trecho a negrito, sem que se verifiquem diferentes interpretações a nível da semântica. Numa das configurações o trecho é considerado uma oração sem verbo, na outra um sintagma nominal. Estas diferentes representações sintácticas têm, além disso, consequências ao nível das funções dos seus constituintes imediatos.

As frases seguintes, por outro lado, são exemplos de ambiguidade ou vagueza[5], com duas leituras possíveis:

CP8-1 A propósito, no Museu da Segunda Guerra Mundial, que aí foi aberto, a história de a maior guerra no continente europeu começa com a fotografia de Estaline a cumprimentar o ministro dos

[5] Para já, a marcação destes casos não faz distinção entre ambiguidade e vagueza. De facto, poder-se-ia argumentar que as duas interpretações nas frases em questão não são contraditórias, podendo mesmo as frases descreverem simultaneamente ambas.

Negócios Estrangeiros da Alemanha **nazi**, ou seja, a guerra começa com a assinatura do Pacto Molotov-Ribbentrop.

CP50-4 Estavam repletas de **lixo, copos de plástico sujos de café**.

CP716-5 José Saldanha acusa a DGA de ter um comportamento «ambíguo e vago» nesta história, «dando cobertura a investigações sem credibilidade nenhuma que foram utilizadas na Bolívia num contexto de luta política pelo poder **nas últimas eleições presidenciais**».

Em CP50-4, *lixo* e *copos de plásticos sujos de café* podem ser considerados como distintos e sintacticamente representados por uma coordenação, ou relacionados, sendo *copos de plástico sujos de café* considerado como uma elaboração de *lixo*, sendo, neste caso, sintacticamente um adjecto predicativo. A ambiguidade presente em CP716-5 refere-se à possibilidade de *nas últimas eleições presidenciais* estar relacionado com dois elementos distintos: *luta política* ou *utilizadas*, que se pauta por uma diferença nos níveis de constituintes, bem como de função sintáctica (adjunto adverbial ADVL vs. dependente nominal N<). No caso de CP8-1, a questão coloca-se igualmente na determinação do que o adjecto *nazi* modifica: *ministro dos negócios estrangeiros da Alemanha* ou apenas *Alemanha*? No entanto, ao contrário dos casos acima, este tipo de ambiguidade apenas implica uma alteração do nível de constituinte de *nazi*, sem alteração da função sintáctica.

De facto, uma das preocupações ancestrais da sintaxe computacional é a ligação dos sintagmas preposicionais, que, a nível sintáctico, apresenta um alto grau de ambiguidade (embora em muitos casos também indicie vagueza – veja-se Santos (1997) e a nota anterior). É interessante referir que essa questão foi também das que denunciou um menor consenso entre anotadores.

Todo o trabalho intelectual tem uma margem de erro e uma margem de arbitrariedade pela subjectividade de interpretação que pressupõe; estamos, portanto, perfeitamente conscientes que há sempre imperfeições e inconsistências em material tão vasto produzido por alteração directa, submetido a um escrutínio em série. Tentámos, contudo, medir de alguma forma estes parâmetros, quer através de um teste inter-anotadores (Afonso, 2001) como de uma análise aprofundada de uma pós-revisão (Afonso, 2004b), para os quais remetemos os leitores interessados.

3. Avaliação conjunta usando a Floresta

A principal razão do presente artigo é sugerir que a Floresta é um óptimo ponto de partida para organizar uma (família de) avaliações conjuntas à volta da sintaxe do português, assim como é um recurso de eleição para ajudar noutras avaliações conjuntas, devido à riqueza da análise e ao cuidado posto na sua criação.

Em primeiro lugar, descreveremos aqui algumas actividades de avaliação feitas internacionalmente usando uma floresta como centro (ou matéria prima).

3.1 Avaliações usando florestas

Em 1992, Grishman e seus colegas (Grishman *et al.*, 1992) congratulavam-se com a existência (de uma versão preliminar) do Penn Treebank (PTB) para poderem aplicar o que foi a primeira métrica desenvolvida em conjunto para análise sintáctica. Oito grupos diferentes, comparando as saídas dos seus sistemas com a análise "ideal" da floresta, concordaram num processo justo de comparação, chamado Parseval (Black *et al.*, 1991), envolvendo os seguintes ingredientes:

- ignorar nomes de nós
- remover uma série de constituintes "leves", tais como auxiliares, pontuação, marcador de infinitivo, etc.
- simplificar parênteses (não aceitando constituintes unários nem vazios)
- calcular o número de chavetas ("brackets") concordante (e o número de cruzamentos ("crossings"), que mede, portanto, a discordância)

No caso de haver mais do que uma análise produzida pelo sistema, convencionou-se usar a primeira. Grishman *et al.* (1992) descreve a avaliação de várias versões de um analisador

usando 317 frases da dita floresta, que se tornou, nas palavras de Gaizauskas *et al.* (1998x), o paradigma dominante (Parseval + Penn Treebank).

Lin (1995), contudo, apontou várias fraquezas deste método, secundado, aliás, por Carroll *et al.* (1998) e Gaizauskas *et al.* (1998a,b):

- o prejuízo de sistemas mais finos (na medida de precisão) – o que levou, aliás, à sugestão de novas medidas;
- a possibilidade de um analisador sintáctico "disparatado" poder obter uma boa classificação com este método (ou seja, a relação entre o número de parênteses concordantes não implica, à partida, quaisquer análises motivadas linguisticamente, veja-se também Sampson & Babarczy, 2003);
- a demasiada dependência de uma dada teoria ou formalismo – em particular, a ligação à gramática generativa – que resulta no "encorajar a produção de analisadores sintácticos e gramáticas [que são] sintonizados pela gramática empregue no corpus de referência" (Gaizauskas *et al.*, 1998x:144, tradução nossa).

De facto, conforme Ringger *et al.* (2004) apontam, não houve muitas avaliações de analisadores sintácticos não ligados umbilicalmente ao Penn Treebank por usarem o mesmo formalismo: Riezler *et al.* (2002), contudo, traduziram o PTB para uma versão LFG-leve, compatível por sua vez com um conjunto de análises LFG mais finas; enquanto Magerman (1995) treinou um analisador sintáctico estatístico, baseado em árvores de decisão, com o material do PTB (traduzido inicialmente para esse formalismo), para depois comparar com a anotação canónica dessa floresta. Neste caso particular, a avaliação também entrou em conta com os nomes dos constituintes, o que podemos considerar um método "Parseval forte". Finalmente, Ringger *et al.* (2004) avaliam um analisador sintagmático baseado noutros princípios usando o PTB.

Em paralelo, houve várias sugestões de avaliação de relações de dependência, consideradas mais semânticas, não só no âmbito da gramática dependencial mas também propondo a transformação dos modelos sintagmáticos para dependenciais, no âmbito da própria gramática generativa. Carroll *et al.* (1998, 2003) propuseram o chamado esquema RG (relações gramaticais – em inglês, "GR schema") e criaram um recurso de avaliação a partir de 500 frases do corpus SUSANNE, em três géneros distintos, revisto manualmente após aplicação de um processo automático. Essa floresta é pública, juntamente com um sistema avaliador, **greval**, baseado no esquema proposto.

3.2 Passos iniciais de reconversão da Floresta

Como autores do recurso, estamos claramente conscientes de que são necessárias algumas tarefas preliminares para converter a Floresta (ou parte dela) num recurso consensual. Não só combinando um (sub)conjunto de categorias sobre cuja definição haja consenso, como eventualmente explicitando informação até agora apenas codificada implicitamente e, muito possivelmente, simplificando as etiquetas até um "mínimo múltiplo comum".

Por exemplo: adverbiais livres, adverbiais associados ao objecto, objectos preposicionais e predicados adverbiais podiam ser amalgamados numa supercategoria "adverbial", ADVL; predicativos pós-nominais e apostos podiam ser juntos numa categoria geral "dependente", DEP, e por aí adiante. Além disso, o problema de diferentes terminologias seria resolvido pela criação de diferentes "zebras", como nas Morfolimpíadas (Costa *et al.*, neste volume).

Esta ideia não é, obviamente, nova. Gaizauskas *et al.* (1998b) propuseram a criação rápida de florestas para avaliação a partir de corpora anotados (ou florestas) mais complexas, sugerindo alguns passos concretos nessa tarefa de simplificação. Estes autores sugerem, contudo, desprezar os nomes dos constituintes, como no Parseval.

Na nossa opinião, e devido ao facto de a existência de terminologia pessoal, não padronizada, ser costume na área da gramática (computacional e não só), seria absolutamente essencial (e benéfico para a Floresta como recurso mais universal, e, portanto, menos idiossincrático) que os vários participantes, numa espécie de ensaio ou preparação de uma

avaliação conjunta, verificassem e validassem partes da Floresta, de forma a estarem claros – e documentados – os limites de concordabilidade e corrrigidos os erros possíveis.

Está claro que haverá sempre um resíduo de incompatibilidade teórica, como foi verificado nas próprias Morfolimpíadas. Nesse aspecto, deviam ser identificadas o que podemos chamar "categorias fracas", ou seja, aquelas em que o grau de concordância entre anotadores é muito baixo. A nossa experiência com o próprio trabalho de anotação e criação da Floresta leva-nos a sugerir que não sejam tomadas em consideração na avaliação comparativa de sistemas, muito embora queiramos salientar o valor da Floresta para precisamente a avaliação de pormenores complexos, eventualmente apenas tratados por um sistema ou dois.

4. Chamada às armas para participar na construção da Floresta

O trabalho investido na criação deste recurso é considerável; assim como na sua documentação, e não menos na sua disponibilização. Contudo, a comunicação entre os utilizadores (actuais mas também potenciais, ou futuros) e os criadores tem sido mínima. Ao fazer uma floresta pública (e não um recurso para aumentar a competitividade relativa dos criadores), seria de esperar que a comunidade envolvida na sintaxe computacional do português contribuísse de várias maneiras: afinal, faz sentido saber o que é que os utilizadores potenciais de um recurso fariam com ele, se o tivessem, para poder desenvolver esse recurso da melhor maneira possível.

Uma questão óbvia é a do género do texto: a Floresta Sintá(c)tica de momento apenas cobre texto jornalístico dos anos 90. Outros géneros, assim como outros períodos da língua, e mesmo transcrições de oral, poderiam vir a ser adicionados.

Além disso, há categorias ou distinções que podem ser extremamente relevantes e que não foram incluídas, ou porque simplesmente não pensámos nisso, ou porque não temos a mesma concepção da importância relativa de diversos fenómenos. Em particular, a nível do discurso, não foi dada qualquer atenção a fenómenos de co-referência ou de estrutura argumentativa, que podem ser fulcrais para outros investigadores.

Por outro lado, e embora a esse nível tenhamos tido alguns pedidos e mesmo participação na criação de novos filtros, gostaríamos que ainda mais interessados propusessem ou criassem filtros ou tradutores para outros formalismos ou necessidades de processamento.

Mas, mais do que tudo, gostaríamos de poder contar com mais grupos que abraçassem o projecto Floresta Sintáctica como seu, e que participassem também na sua criação e revisão, quer através de trabalho linguístico de revisão quer através da validação ou do fornecimento de material produzido com os seus próprios analisadores. Porque, muito mais difícil de resolver do que o problema dos dados esparsos, é o problema dos linguistas computacionais esparsos!

E parece-nos que, mesmo parcelarmente, a riqueza da informação contida na Floresta Sintá(c)tica justifica o seu uso em avaliações conjuntas mais específicas, como é o caso de futuras morfolimpíadas ou do reconhecimento de entidades mencionadas, para citar apenas algumas das áreas cobertas por este livro.

Pós-escrito: Apraz-nos verificar que, dois anos após a escrita do presente artigo, as sugestões que aqui fizemos quanto ao uso da Floresta em avaliações conjuntas se tornaram realidade: não só material da Floresta foi usado na construção do recurso dourado do HAREM, como a própria Floresta (ou melhor, o Bosque) é neste momento empregue numa avaliação conjunta internacional, a CoNLL-X de 2006. Finalmente, o leque de utilizadores também se estendeu a um público não lusófono, como o demonstra o uso da Floresta no treino de analisadores sintácticos probabilísticos pela Universidade do Texas em Austin (Baldridge & Wing, 2006).