Floresta Sintá(c)tica: um "treebank" para o português

Susana Afonso
Eckhard Bick
Projecto VISL
Universidade do Sul da Dinamarca
Renato Haber
Diana Santos
Processamento computacional do português
SINTEF Telecom & Informatics

1. Introdução: motivação e objectivos

A Floresta Sintá(c)tica compõe-se de um conjunto de textos – frases – sintacticamente analisados, em forma de árvore, previamente revistas intelectualmente. De forma a serem usadas por uma comunidade mais vasta do que os próprios compiladores apenas, eventualmente para efeitos de avaliação conjunta, as árvores foram sendo tornadas publicamente acessíveis na rede[1].

Um dos objectivos da criação de um "treebank" para português era congregar todos os interessados na análise computacional do português, de forma a que a Floresta Sintá(c)tica pudesse reflectir um consenso entre as possibilidades de análise, ou pelo menos permitir uma escolha informada. Assim, uma das esperanças acalentadas pelo presente projecto era a de que este desse origem à discussão e cooperação entre os vários actores, além da criação dos próprios objectos (árvores) e da obtenção de documentação que reflicta progresso em sintaxe computacional da língua portuguesa. Tal ainda não se verificou, talvez por falta de disseminação da própria existência do projecto, falha essa que este artigo pretende (parcialmente) colmatar.

Subjacente ao projecto está a noção de que a existência de recursos linguísticos partilhados por uma comunidade que processa uma dada língua é fundamental, e que o progresso numa dada área exige a comparação de resultados entre grupos diferentes. De facto, é cada vez mais universalmente reconhecida a possibilidade de avaliação de um dado projecto (baseada em recursos públicos) como um *sine qua non* para uma investigação responsável (cf. Gaizauskas, 1998 e Hirschman, 1998).

No campo da linguística computacional, a anotação da estrutura sintáctica de um corpus torna explícita uma quantidade muito maior de informação que permite aplicações computacionais muito mais complexas. Corpora anotados sintacticamen-

[1] Nos endereços http://cgi.portugues.mct.pt/treebank/PaginaFloresta.html e http://visl.sdu.dk/visl/pt/treebank.html.

Actas do XVII Encontro Nacional da Associação Portuguesa de Linguística, Lisboa, APL, 2002, pp. 533-545

te começam a ser uma realidade para várias línguas, e não quisemos que o português ficasse para trás.

2. Organização

Este projecto foi concebido, na sua fase inicial, como um projecto de colaboração entre dois grupos, ambos com experiência na anotação e processamento de corpora anotados e já com um passado de colaboração prática evidenciado pelo projecto AC/DC (Santos & Bick, 2000).

O projecto VISL é um projecto de pesquisa e ensino na Universidade do Sul da Dinamarca iniciado em 1996, baseado em análise computacional automática. Partindo do sistema português PALAVRAS (Bick, 2000) como modelo para outras línguas, a equipa do VISL construiu um núcleo de ferramentas e bancos de dados linguísticos para usar através da rede (WWW). Trabalha-se hoje com a gramática e especificamente a sintaxe de catorze línguas, seis das quais com análise automática segundo o paradigma da "Constraint Grammar" (CG), de Karlsson et al. (1995). Áreas mais recentes de actividade são a semântica e a tradução automática e a recolha e etiquetagem de corpora. Além da possibilidade de interrogação livre através da rede, foi estabelecida uma base de orações controladas para todas as línguas VISL, cobrindo vários fenómenos sintácticos de uma maneira mais sistemática.

Para a aplicação do ensino apoiado por computador, os utilizadores podem escolher entre diversos filtros notacionais, correspondendo a diferentes paradigmas descritivos da língua. Por exemplo, essa interface apresenta exercícios nos quais as palavras são coloridas conforme a classe morfossintáctica a que pertencem, assim como permite ao estudante construir árvores sintácticas gráficas, com etiquetas de forma e função em cada nó, depois controladas automaticamente pelo computador.

O projecto Processamento computacional do português (Santos, 2000), que evoluiu recentemente para um centro de recursos distribuído para o processamento da língua portuguesa, é um projecto lançado pelo Ministério da Ciência e da Tecnologia para melhorar o estado da área, considerada prioritária. Um dos seus principais métodos de actuação é a criação de recursos públicos para a investigação e desenvolvimento na área do processamento computacional do português, tendo lançado (em alguns casos em colaboração) vários projectos de disponibilização de recursos, tal como o AC/DC, o COMPARA, o CETEMPúblico e a própria Floresta Sintá(c)tica. Outra das prioridades deste projecto/centro é a avaliação.

Do ponto de vista do projecto VISL, teria interesse aumentar significativamente o conjunto de frases analisadas e revistas intelectualmente para aumentar as capacidades de ensino do sistema, além de, em geral, permitir a melhoria do analisador sintáctico PALAVRAS subjacente. A principal motivação para o projecto Processamento computacional do português participar na Floresta era a construção de um recurso que pudesse eventualmente ser usado para a avaliação de analisadores sintácticos e outras ferramentas computacionais, a partir de uma base de objectos (árvores) comum, validada por linguistas.

Embora a motivação dos dois grupos fosse distinta, considerámos tal característica enriquecedora, dado que a satisfação de ambos os objectivos era realizável simultaneamente, embora talvez de forma mais demorada. Pensou-se também que a existência de uma primeira fase de experimentação e definição mais rigorosa das especificações seria útil antes de lançar um projecto colaborativo muito maior, abrangendo virtualmente todos os grupos envolvidos em análise sintáctica automática.

O material base para um "treebank" teria necessariamente que ter o problema dos direitos de autor resolvido, por isso decidiu-se usar o primeiro milhão de palavras do CETEMPúblico (Rocha & Santos, 2000) e envidar esforços na obtenção de material semelhante para o português brasileiro.

3. Outros projectos

Existem e existiram vários projectos[2] cujo objectivo é criar um conjunto de objectos linguísticos que sirvam como recurso para várias tarefas em engenharia de linguagem e linguística. Contudo, existem grandes diferenças sobre a forma de prosseguir e mesmo sobre a própria definição de "treebank".

Desde o Penn Treebank (Marcus et al., 1993) e o corpus SUSANNE (Sampson, s.d.) para o inglês e o Prague Dependency Treebank (Hajic, 1998) para o checo, precursores do moderno emaranhado de florestas, e seguindo simplesmente um dado formalismo linguístico (sintagmático no caso do Penn, dependencial no caso do PDT), até ao TIGER (Dipper et al., 2001) para o alemão, cujo formalismo sofisticado junta propriedades destes dois tipos de representação linguística, muitas variantes se podem encontrar, a que não teremos infelizmente ocasião de fazer referência detalhada aqui.

Plantação da Floresta: descrição do processo

Podemos considerar duas fases distintas na construção da Floresta Sintá(c)tica: uma fase de pré-processamento e a fase de revisão da análise automática propriamente dita (tanto no formato CG como no formato de árvores gerado a partir deste).

A fase de pré-processamento reviu aspectos que não são morfossintácticos nem estruturais. Quanto à parte lexicográfica, que consistiu no enriquecimento do léxico português do PALAVRAS com a inserção das palavras mais frequentes do corpus (cerca de 8.000 a 9.000 novos lexemas), esta revelou-se de extrema utilidade, evitando erros de análise automática devido a falhas no dicionário do PALAVRAS.

A fase de pré-processamento englobou também uma revisão manual da separação frásica automática. Esta tarefa acabou por ser desenvolvida num espaço de tempo bastante mais longo do que seria desejado. A razão está em que o que ini-

[2] Mais de uma dezena de exemplos de diferentes treebanks/florestas são mencionados pelo projecto TIGER, em http://www.ims.uni-stuttgart.de/projekte/TIGER/related/links.shtml.

cialmente se previa ser uma simples revisão da separação automática, acabou por ser, de facto, um questionamento dos critérios automáticos de separação frásica (porque baseada em critérios exclusivamente de pontuação e de ortografia), o que conduziu à redefinição dos critérios de frase e divisão frásica (Afonso & Marchi, 2001a).

Esta redefinição impulsionou uma outra questão relacionada com as frases que seriam revistas. Decidiu-se que certas estruturas complexas, como versos de poemas, ou estruturas não consideradas como interessantes a serem analisadas sintacticamente (como resultados desportivos) fossem etiquetadas com <sic> e não analisadas (Afonso & Marchi, 2001b).

A segunda fase corresponde à criação automática e revisão manual do corpus em formato CG, seguido pela geração automática e revisão manual de árvores deitadas. Dado o grande número de categorias e distinções já usado pelo analisador automático PALAVRAS, e o interesse dos participantes em criar um corpus para uso o mais variado possível, optou-se por fazer uma revisão morfossintáctica exaustiva, a nível da função e forma sintáctica e informação morfológica.

Uma vez que a análise automática progride em dois passos separados (CG e geração de árvores), optou-se por também dividir o trabalho de revisão em duas etapas: primeiro, a revisão de etiquetas CG de forma e função e de constituintes sintácticos tipo estrutura sintagmática ("phrase structure grammar") (árvores deitadas). Desta maneira, evita-se que certos erros sejam propagados pela geração automática de árvores. Por exemplo, um erro em termos de função CG na fase 1, como um sujeito num lugar errado, pode inibir a criação duma oração constituinte na fase 2, porque as regras gramaticais do gerador sintáctico não conseguem colocar esse sujeito extra numa estrutura que faça sentido. Aqui, uma correcção na fase 1 evita muitas intervenções manuais na fase 2, aumentando a robustez de todo o processo e tornando a revisão humana mais eficaz. A bipartição do processo facilita a manutenção das gramáticas CG e de estrutura sintagmática, permitindo ao autor do PALAVRAS remediar os problemas correspondentes da análise automática no lugar certo.

Além disso, dado que tanto a análise automática como a revisão humana subsequente são feitas em trechos de algumas centenas de frases de cada vez, problemas de análise e criação de árvores identificados num trecho anterior podem ser remediados antes da ronda seguinte de análise. Mais, o tratamento por fases permitiu implementar distinções novas de anotação, não só na floresta, produto final, mas também, até um certo ponto, no sistema automático.

Apesar de, à primeira vista, este trabalho parecer linear, envolve questões de difícil resolução. É fundamental ter em conta que se trata de um corpus de textos jornalísticos produzidos por um dado falante, ou seja, trata-se de um corpus que reflecte a realidade linguística e, como tal, estruturas de difícil representação sintáctica foram frequentes, como hesitações (no caso de transcrição de entrevistas), discurso indirecto livre e mesmo erros de expressão, que se decidiu não corrigir, por forma a introduzir o mínimo de alterações no corpus original. A estrutura dos títulos também se revelou de complexa representação. Neste ponto, o trabalho de anotador

foi também na direcção da identificação destes casos e da resolução dos mesmos, em termos de representação formal. Uma das soluções encontradas foi a introdução manual de códigos específicos para identificação de determinadas estruturas (por exemplo, #D para estruturas discursivas, #E para elipse, etc.).

4.1. Revisão do formato CG (Constraint Grammar)

A revisão das frases em formato CG implica a revisão da informação morfossintáctica que cada lexema possui. A informação morfológica consiste na classe de palavras (categoria gramatical), traços morfológicos, e forma base (lematização). A informação sintáctica, expressa também em cada um dos elementos, consiste na função sintáctica interna a cada uma das orações, e na função externa, ou seja, a função de cada oração subordinada na frase, codificada no verbo principal da oração, ou no complementarizador.

Porque a CG é uma gramática dependencial, marcadores de dependência (<, >) indicam a direcção do núcleo sintáctico de que os constituintes são dependentes (excepto o verbo principal, que não exibe marcadores dependenciais). Veja-se um exemplo concreto de revisão de CG, apresentando primeiro a sua análise automática, e em seguida a análise revista por um linguista (alterações a negrito):

Os [o] <art> DET M P @>N
que [que] <rel> SPEC M S @SUBJ> @#FS-N<
escolheram [escolher] <fmc> V PS/MQP 3P IND VFIN @FMV
sábado [sábado] N M S @<ADVL
podem [poder] <fmc> V PR 3P IND VFIN @FAUX
começar [começar] V INF @IMV @#ICL-AUX<
cedo [cedo] ADV @<ADVL

Os [o] **<dem>** DET M P **@SUBJ>**
que [que] <rel> SPEC M S @SUBJ> @#FS-N<
escolheram [escolher] <fmc> V PS/MQP 3P IND VFIN @FMV
sábado [sábado] N M S **@<ACC**
podem [poder] <fmc> V PR 3P IND VFIN @FAUX
começar [começar] V INF @IMV @#ICL-AUX<
cedo [cedo] ADV @<ADVL

Uma vez que as árvores são geradas a partir do formato CG, é importante que estas categorias sejam revistas de forma cuidadosa. Irregularidades detectadas antes da revisão do formato de árvores diminuirão o tempo gasto na revisão da informação morfossintáctica, nesta segunda etapa,[3] que envolve outras dimensões de revisão.

[3] Não é realista considerar que todas as irregularidades a esse nível sejam detectadas na primeira fase, já que uma margem de erro humano deve ser tida em conta.

4.2. Revisão no formato de árvores deitadas

O formato de árvores deitadas tem as seguintes características: cada nível abaixo do nó mais alto, é indentado, ou seja, o seu nível de constituinte é representado por sinais de igual. A revisão das árvores geradas a partir do formato de CG revisto envolve, desta forma, vários níveis: além da verificação da informação morfossintáctica, a parte crucial da revisão nesta fase é a revisão dos níveis de constituintes.

Se é facilmente deduzível que uma correcção da informação morfossintáctica num estádio prévio ao formato de árvores irá ter consequências directas aquando da geração das árvores, o mesmo já não se pode dizer em relação aos níveis de constituintes, exactamente porque o formalismo da CG é dependencial de superfície. Por exemplo, em estruturas como SN com complementos preposicionais, a dependência relativamente ao substantivo é realizada por uma seta dependencial, normalmente, imediatamente à sua esquerda. Ou seja, se essa dependência for a um substantivo mais afastado, o formalismo de CG não permite essa distinção, sendo a etiqueta de função a mesma, @N<, e a correspondência desta relação em formato de árvores não directa. Vejamos, por exemplo, o seguinte SN, em CG, correspondente ao texto *O empregado de balcão do café Magestic*:

```
O          [o] DET M S @>N
empregado  [empregado] N M S @NPHR
de         [de] PRP @N<
balcão     [balcão] N M S @P<
de         [de] PRP @N<
o          [o] DET M S @>N
café       [café] N M S @P<
Magestic   [Magestic] PROP M S @N<
```

Cognitivamente não parece haver qualquer ambiguidade na análise do sintagma preposicional *do café Magestic*; qualquer falante não terá dúvidas em considerar que é do empregado do café Magestic que aqui se trata, não do balcão do café Magestic. No entanto, o analisador automático não possui mecanismos que permitam desambiguar esta questão e, em termos puramente sintáctico-formais em CG, a relação de dependência referida nem sequer é expressa. O que é representado é apenas a dependência de *de o café Magestic* do núcleo nominal à sua esquerda (tanto pode ser balcão, como empregado). A geração da árvore correspondente não vai tomar o substantivo mais longínquo como o núcleo do sintagma de que o SP é dependente, mas aquele que imediatamente o precede: balcão.

```
UTT:np
>N:art('o' <artd> M S)  O
H:n('empregado' M S)    empregado
N<:pp
```

```
=H:prp('de')  de
=P<:np
==H:n('balcão' M S)  balcão
==N<:pp
===H:prp('de' <sam->)  de
===P<:np
====>N:art('o' <-sam> M S)  o
====H:n('café' M S)  café
====N<:prop('Magestic' M S)  Magestic
```

Alguns casos de dependência podem ser solucionados em termos formais, mas casos outros há, como o do exemplo anterior, que necessitam de desambiguação humana. De facto, um dos critérios de "plantação" de árvores foi considerar análises que reflectissem apenas a interpretação humana (desfazendo-se as ambiguidades) e não todas as análises possíveis em termos puramente sintácticos.[4] No entanto, em casos de real ambiguidade, foi especificada uma notação exprimindo duas ou mais análises, na mesma frase ou, se impossível desta forma, em frases diferentes (A1, A2, etc.) (cf. Afonso et al., 2001).

Estes são exemplos simples de revisão, que implicam apenas a diminuição / aumento da indentação, criação / eliminação de nós constituintes, tarefas para as quais foi desenvolvida a ferramenta Pica-pau. Saliente-se, contudo, que a equipa de anotadores da Floresta contou sempre com a visualização das árvores na sua forma gráfica através do sítio do VISL, recurso este que se revelou de extrema utilidade, dado que é mais fácil detectar irregularidades nesta forma do que nas árvores deitadas.

De qualquer maneira, a equipa da Floresta foi confrontada com problemas de representação formal complexos, tendo de tomar opções linguísticas de base para os colmatar. Um dos casos complexos foram as estruturas elípticas. Todas as opções linguísticas tomadas no âmbito da Floresta têm de respeitar os princípios notacionais e de terminologia que regem o projecto VISL. Ou seja, tem de haver um compromisso entre a notação e princípios já estabelecidos e opções de análise linguística. Neste contexto, e referindo-nos às estruturas elípticas, a solução encontrada para as representar respeitou o princípio de não existência de constituintes nulos. Significa isto que o elemento elíptico não poderia ser representado por um constituinte vazio (Ø). Numa frase como por exemplo *Os quatro primeiros temas destinam-se a mostrar o papel de Portugal no mundo e o quinto é justificado por a experiência de Barcelona (Port Aventura)*, não se pôde considerar *tema* como constituinte vazio, o que simplificaria a análise:

[4] Outro factor é o conhecimento extra-linguístico. Embora válido em termos sintácticos, a oração relativa na frase *Em relação ao Iraque, Valeri Progrebenkov (...) desmentiu a existência de uma encomenda de 4000 carros de combate russos, como afirmara o genro de Saddam Hussein* ***que desertou para a Jordânia.*** *(...)* não se pode referir a Saddam Hussein...

```
STA:cu
CJT:fcl
=SUBJ:np
==>N:art('o' M P)  Os
==>N:num('quatro' <card> M P)  quatro
==>N:adj('primeiro' <NUM-ord> M P) primeiros
==H:n('tema' M P)    temas
=P:v-fin('destinar' PR 3P IND)   destinam-
=ACC:pron-pers('se' M 3P ACC)      se
=PIV:pp a mostrar o papel de Portugal em o mundo
CO:conj-c('e' <co-subj>)  e
CJT:fcl
=SUBJ:np
==>N:art('o' M S)  o
==>N:adj('quinto' <NUM-ord> M S)  quinto
==H:ø      Ø
=P:vp      é justificado
=PASS:pp   por a experiência de Port-Aventura (Barcelona)
```

Outro tipo de representação teve, pois, de ser convencionada. Em primeiro lugar, houve a necessidade de se definir elipse em traços largos. Concluiu-se que eram considerados casos de elipse elementos cuja reconstrução seria possível pelo contexto frásico. A partir deste princípio de "recuperabilidade", optou-se por representar estas estruturas, mantendo, por assim dizer, o elemento elíptico visível ao preservar a função sintáctica que os constituintes apresentariam se o elemento não fosse elíptico. No caso acima, esta solução apresentar-se-ia desta forma:

```
STA:cu
CJT:fcl
=SUBJ:np
==>N:art('o' M P)  Os
==>N:num('quatro' <card> M P)  quatro
==>N:adj('primeiro' <NUM-ord> M P) primeiros
==H:n('tema' M P)    temas
=P:v-fin('destinar' PR 3P IND)   destinam-
=ACC:pron-pers('se' M 3P ACC)      se
=PIV:pp    a mostrar o papel de Portugal em o mundo
CO:conj-c('e' <co-subj>)  e
CJT:fcl
=SUBJ:np
==>N:art('o' M S)  o
==>N:adj('quinto' <NUM-ord> M S)  quinto
=P:vp          é justificado
=PASS:pp       por a experiência de Port-Aventura (Barcelona)
```

No entanto, se um nó constituinte é constituído por um determinante ou outro modificador e um núcleo elíptico, a análise descrita não pode ser aplicada, porque colide com outro princípio básico: a não existência de nós com um só membro / dependente. Desse modo, o determinante terá de passar a núcleo. Veja-se, por exemplo, a frase *Comem dois pães ao pequeno-almoço e três Ø ao lanche*. Se se mantiver a função sintáctica de *três* (@>N) e estando o núcleo elíptico (*pães*), teríamos um grupo constituído apenas pelo modificador nominal. Por isso, para estes casos, o numeral passa a ser o núcleo do SN, o que é, aliás, a estratégia mais conforme à CG.

Para facilitar a busca destes casos, estabeleceram-se códigos #E para todos os casos de elipse, subdividida depois nos seus tipos (elipse de grupo <Eg>, elipse sintáctica <Es> e elipse morfológica <Em>).

5. Ferramentas desenvolvidas

Durante o primeiro ano do projecto foram desenvolvidas duas ferramentas, infelizmente coincidindo o tempo da sua especificação e desenvolvimento com o próprio trabalho de construção da floresta, o que levou a que não fossem quase usadas durante a construção do recurso. Contudo, poderão ser úteis quer para a exploração do resultado por utilizadores externos (o Águia), quer em futuras fases de criação de novas árvores revistas (o Pica-pau).

O Pica-pau pretende facilitar o trabalho de edição das árvores sintácticas, permitindo deslocar nós-terminais (palavras e sinais de pontuação) e nós constituintes, bem como criar novos nós constituintes sem que o utilizador precise de verificar a indentação dos nós a serem deslocados ou inseridos. Esta ferramenta actua justamente na questão de manter a estrutura da árvore. Para mais informações sobre cada comando, bem como um exemplo detalhado de utilização da ferramenta, veja-se Haber (2001).

O objectivo primordial do Águia é permitir uma procura global em corpora, baseada em atributos sintácticos e não lexicais, através da rede. O Águia foi concebido para permitir duas actividades diferentes: em primeiro lugar, a visualização / interrogação do resultado da anotação e revisão da Floresta por todos os interessados; em segundo lugar, a determinação e identificação, por parte da própria equipa da Floresta, de problemas na anotação automática que possam ser resolvidos sistematicamente (em futuras versões do analisador sintáctico) sem recurso a uma modificação manual repetitiva. Esta ferramenta constitui não só uma extensão natural ao sistema de procura do projecto AC/DC, em que os elementos básicos são as palavras, única entidade que apresenta classificação, mas também uma extensão natural à interface do projecto VISL, em que se pode inspeccionar **cada** árvore individualmente mas não em conjunto.

6. Resultado do projecto

Durante a sua primeira fase (correspondendo a aproximadamente um ano de trabalho), o projecto Floresta Sintáctica produziu o *Bosque*: 1.427 árvores (correspondendo a 251 extractos, 1.405 frases distintas, 36.408 unidades, aprox. 34.256 palavras) revistas e a *Floresta Virgem*: 41.406 árvores, ou seja, o primeiro milhão de palavras do CETEMPúblico analisado e arborizado automaticamente, sem revisão (7.913 extractos, 41.406 frases, 1.072.857 unidades).

Cada árvore da nossa Floresta corresponde a três objectos diferentes: uma análise sintáctica dependencial (formato CG), por palavra; uma análise sintáctica de constituintes (formato árvores deitadas, ficheiro de texto); e uma análise sintáctica de constituintes (formato árvores gráficas, figura).

Outro dos resultados do projecto, sem o qual os objectos mencionados não seriam interpretáveis, é a documentação associada. A documentação num projecto desta natureza é fundamental, por várias razões. Em primeiro lugar, porque a informação que envolve é muito vasta, e é preciso produzir diferentes tipos para diferentes usos, desde a informação nos sítios WWW do projecto, à identificação e documentação de todas as etiquetas formais e meta-anotação (códigos) utilizadas na análise automática das frases[5], passando pela especificação formal do formato das frases intelectualmente revistas[6] até à descrição das opções linguísticas tomadas durante o processo. Apenas deste modo a Floresta Sintá(c)tica se torna legível e compreensível para o utilizador.

Do ponto de vista do anotador, documentar opções linguísticas significa uma reflexão profunda sobre o tipo de problemas que uma frase pode levantar em termos de análise e representação formal. Através de um trabalho de reflexão, mais facilmente se atinge uma maior consistência no corpus revisto intelectualmente, especialmente quando se trata de mais de um anotador (caso deste projecto).

A documentação linguística encontra-se dividida em duas partes distintas: uma referente a opções genéricas, ou seja, de carácter formal que ultrapassam o projecto da Floresta Sintá(c)tica, uma vez que são princípios que regem o próprio projecto VISL. São linhas gerais que presidem à anotação. Outra parte relaciona-se com opções de carácter linguístico que presidem à revisão da análise automática que surgiram da necessidade sentida em definir linhas de orientação de forma a fazer face a problemas de representação formal.

7. Teste inter-anotadores

A realização de um teste inter-anotadores impõe-se em qualquer projecto em que a consistência das análises revistas pelo mesmo ou por mais do que um anotador deve ser a mais elevada possível. No caso da Floresta Sintá(c)tica, o teste foi

[5] Presente no Glossário: http://cgi.portugues.mct.pt/treebank/glossario.html

[6] Em http://cgi.portugues.mct.pt/treebank/BNFfloresta.html.

elaborado com os objectivos de listar situações frequentes de desacordo entre anotadores, contar diferenças observadas entre os ficheiros revistos paralelamente por vários anotadores, documentar as diferenças e contabilizar as alterações em relação ao ficheiro automaticamente analisado.

A metodologia adoptada foi a seguinte: três anotadores reviram 107 árvores em paralelo e sem qualquer discussão comum, no prazo de uma semana. A revisão foi realizada directamente no formato de árvores deitadas, sem recurso à sua forma gráfica, por razões práticas, uma vez que era o mesmo conjunto de frases que estaria a ser visualizado simultaneamente, o que implicaria uma sobreposição de análises.

Os três ficheiros revistos foram comparados dois a dois (**R**(evisão)**1** e **R**(evisão)**2**; **R1** e **R3**; **R2** e **R3**) através do comando de Unix/Linux *diff* e o resultado dessa comparação quantificado em categorias estabelecidas para o efeito (diferenças observadas e razões das diferenças observadas) e segundo princípios de contagem definidos.

A quantificação das diferenças foi feita por par, o que significa que, para cada categoria, o número de diferenças era três, no caso de os três anotadores exibirem análises diferentes, dois (no caso de um anotador divergir dos outros dois) ou zero (se todos apresentarem a mesma análise).

Uma versão final da análise das 107 frases foi depois estabelecida, após discussão comum. Para uma descrição completa do processo de realização da experiência de teste inter-anotadores e suas conclusões, consulte-se Afonso (2001).

8. Reflexão para futuro trabalho florestal

Durante todo o processo de construção da Floresta, um trabalho necessário de reflexão foi sendo feito, avaliando-se o processo à medida que este ia decorrendo: opções tomadas, resultados alcançados, trabalho de revisão. Note-se que este é o primeiro projecto deste tipo para o português e, como tal, este primeiro ano da Floresta Sintá(c)tica foi um ano também de experimentação, de discussão de possibilidades.

Olhando para os resultados directos da Floresta Sintá(c)tica, foram revistas 1427 frases, ou seja, cerca de 10% do primeiro milhão do CETEMPúblico. Apesar de, numa primeira leitura, este resultado não constituir ainda uma floresta, estes 10% correspondem a uma revisão exaustiva das frases a todos os níveis. Além disso, este conjunto de frases levou a todo um trabalho de recolha de tipos de problemas, bem como à discussão, implementação e documentação de opções linguísticas tomadas para os resolver.

Desta forma, considera-se este trabalho inicial de extrema importância para todo o processo de revisão, uma vez que ao estudarem-se possibilidades de análise, critérios de "plantação" de árvores e resolução de casos problema, prepara-se uma futura revisão do corpus de forma mais sólida / consistente. Além disso, fruto deste

trabalho inicial, o analisador automático foi-se adaptando também às novas necessidades, sendo progressivamente melhorado.

Agradecimentos

A equipa de Floresta incluiu a Ana Raquel Marchi, que lamentamos não ter podido participar na escrita do presente artigo. Estamos gratos a Mogens Svendsen pelo apoio prestado na confecção do poster apresentado à conferência.

Referências

Afonso, Susana. "Na trilha de um teste inter-anotadores", 2001, http://cgi.portugues.mct.pt/treebank/TrilhaTIA.rtf.

Afonso, Susana e Ana Raquel Marchi. "Critérios de separação de sentenças/frases", 2001a, http://cgi.portugues.mct.pt/treebank/CriteriosSeparacao.html

Afonso, Susana e Ana Raquel Marchi. "A etiqueta <sic> </sic>", 2001b. http://cgi.portugues.mct.pt/treebank/CriteriosSic.html

Afonso, Susana, Eckhard Bick e Ana Raquel Marchi. "Notational and terminological guide-lines", 2001, http://www.visl.hum.sdu.dk/visl/pt/guidelines.html

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

Gaizauskas, Robert. "Evaluation in language and speech technology". *Computer Speech and Language*, 12 (4) (1998), pp.249-62.

Dipper, Stefanie, Thorsten Brants, Wolfgang Lezius, Oliver Plaehn & George Smith. "The TIGER treebank", apresentado ao *Third Workshop on Linguistically Interpreted Corpora*, www.ims.uni-stuttgart.de/projekte/TIGER/paper/linc2001-abstract-tiger.pdf.

Hirschman, Lynette. "The evolution of Evaluation: Lessons from the Message Understanding Conferences", *Computer Speech and Language* 12 (4) (1998), 281-305.

Haber, Renato Ribeiro. "Pica-pau: Um protótipo de ferramenta para visualização e edição de árvores sintáticas", http://cgi.portugues.mct.pt/treebank/Picapau.html.

Hajič, Jan. "Building a Syntactically Annotated Corpus: The Prague Dependency Treebank", in *Issues of Valency and Meaning*, Karolinum, Praha 1998, pp. 106-132.

Karlsson, F., A. Voutilainen, J. Heikkilä & A. Anttila. *Constraint Grammar: ALanguage-Independent Framework for Parsing Unrestricted Text*. Mouton de Gruyter, Berlin / New York, 1995.

Marcus, Mitchell P., Beatrice Santorini & Mary Ann Marcinkiewicz. "Building a large Annotated Corpus of English: The Penn Treebank", *Computational Linguistics* **19** (2), June 1993, 313-30.

Rocha, Paulo & Diana Santos. "CETEMPúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa", in Maria das Graças Volpe Nunes (ed.), *Actas do V Encontro para o processamento computacional da língua portuguesa escrita e falada, PROPOR'2000* (Atibaia, SP, Brasil, 19-22 de Novembro de 2000), 131-140.

Sampson, Geoffrey. "SUSANNE Corpus and Analytic Scheme", http://www.cogs.susx.ac.uk/users/geoffs/RSue.html.

Santos, Diana. "O projecto Processamento Computacional do Português: Balanço e perspectivas", in Maria das Graças Volpe Nunes (ed.), *Actas do V Encontro para o processamento computacional da língua portuguesa escrita e falada, PROPOR'2000* (Atibaia, São Paulo, Brasil, 19-22 de Novembro de 2000), 105-113.

Santos, Diana & Eckhard Bick. "Providing Internet access to Portuguese corpora: the AC/DC project", in Gavriladou et al. (eds.), *Proc. Second International Conference on Language Resources and Evaluation, LREC2000* (Athens, 31 May-2 June 2000), 205-10.