Floresta Sintá(c)tica: primeiro ano

Susana Afonso
Eckhard Bick
Projecto VISL
Universidade do Sul da Dinamarca

Renato Haber
Diana Santos
Processamento computacional do português
SINTEF Telecom & Informatics

Versão **1.0**, 13 de Fevereiro de 2002
http://cgi.portugues.mct.pt/treebank/Afonsoetal2002.rtf
http://cgi.portugues.mct.pt/treebank/Afonsoetal2002.ps
http://cgi.portugues.mct.pt/treebank/Afonsoetal2002.doc

O presente artigo é uma versão alargada do artigo "Floresta sintáctica: um *treebank* para o português", a publicar nas Actas do XVII Encontro da Associação Portuguesa de Linguística (Lisboa, Outubro de 2001).

1 Motivação

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 rede1.

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).

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

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 sintacticamente 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 Proc. Comp. Port. 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 satissfação 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 usar-se o primeiro milhão de palavras do CETEMPúblico (Rocha & Santos, 2000) e envidar esforços na obtenção de materail 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.

4 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

[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.

inicialmente 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 heurísticos, exclusivamente na pontuação e na ortografia), o que conduziu à redifinição dos critérios de frase e divisão frásica (Afonso & Marchi, 2000a). Para um projecto essencialmente sintáctico, a equipa da Floresta considerou que os critérios inicialmente definidos não constituíam uma boa base.

A revisão / redefinição dos critérios de separação frásica 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, 2000b). Analogamente, decidiu-se não analisar fragmentos de frase (<s frag>) ou autores <a>, mas optou-se por analisar listas (<li>), por apresentarem uma estrutura oracional. Os títulos foram também analisados e alguns extractos foram eliminados do corpus.

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 das frases que compõem o Bosque.

Dado o grande número de categorias e distinções já usado pelo analisador automático PALAVRAS, e o interesse linguístico 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.

No que respeita directamente à especificação versus não especificação, a equipa da Floresta acabou por concluir que, em determinados casos, seria preferível a não especificação, porque essa decisão pode não ser mantida de forma consistente para todo o corpus. Tal é o caso, por exemplo, da especificação do género de topónimos, uma vez que diversas interpretações podem surgir quando se questiona o género dos topónimos que em português não se fazem acompanhar de artigos. No caso de Aveiro, poder-se-ia chegar a mais do que uma conclusão: porque aparenta ser uma palavra do género masculino, o género seria masculino, porque é uma cidade (a cidade de Aveiro), seria feminino.

Dado 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 "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 correção na fase 1 evita muitas intervenções manuais na fase 2, aumentando a robustez de todo o processo e tornando a revisão humana mais eficaz. Além disso, 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, em dado grau, no sistema automático.

Optou-se por dividir o trabalho de revisão em duas etapas: primeiro a revisão das frases em CG e posteriormente no formato de árvores deitadas, geradas a partir do formato CG modificado. A razão da existência destas etapas relaciona-se com a convicção de que, desta forma, o sistema de revisão seria mais robusto, uma vez que o anotador, não tendo de rever níveis de constituintes – porque a CG é um formalismo plano (sem níveis) –, limita a revisão à informação morfossintáctica. Uma vez que a geração das árvores deitadas é realizada a partir do formato CG revisto, a revisão das árvores deitadas significa, prioritariamente, uma revisão dos níveis de constituintes.

Apesar de, à primeira vista, este trabalho parecer linear, envolve questões de difícil resolução. É fundamental ter 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 representação3. 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 específicos4 para identificação de determinadas estruturas (por exemplo, #D para estruturas discursivas, #E para elipse).

Apresenta-se de seguida cada uma das etapas da fase de revisão mais pormenorizadamente.

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), género e número, em particular, para substantivos, adjectivos e particípios passados; tempo, modo e pessoa para verbos e caso para pronomes, além do resultado da sua lematização, a sua forma base.

A informação sintáctica está também expressa em cada um dos elementos, tanto a que respeita à função sintáctica interna a cada uma das orações, bem como a 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).

[3] A descrição da estrutura / sintaxe dos títulos ultrapassa o âmbito deste artigo e por isso, este tema, vasto, não será aqui discutido.

[4] A lista de códigos definida pode ser consultada em http://visl.hum.sdu.dk/visl/pt/MainDocumentationTreebank.html, ponto A.

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,[5] que envolve outras dimensões de revisão.

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.[6] 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

[5] 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, mas é de esperar que o número dessas irregularidades a nível puramente morfossintáctico seja consideravelmente reduzido e que, por esse motivo, serão poucos os casos em que se produzirão alterações, nesta matéria, no formato de árvores.

[6] A excepção é o nível imediatamente abaixo do nó mais alto, por se considerar que as estruturas são mais facilmente manipuláveis se não for introduzido um sinal de igual a este nível; seria, de certa forma redundante indentar o nível que segue a raiz da árvore, já que que todas as árvores possuem uma raiz.

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[7], 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. Em termos de indentação, isto significa que o SP vai estar ao mesmo nível (vai exibir o mesmo número de sinais de igual) que o núcleo do SN, 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

[7] Ou, ((o empregado (de balcão)) do café Magestic), núcleo complexo.

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.[8]

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., 2000).

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, abaixo descrita. 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:

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

[8] 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...

=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(M P) Os
==>N:num(<card> M P) quatro
==>N:adj(<NUM-ord> M P) primeiros
==H:n(M P) temas
=P:v-fin(PR 3P IND) destinam-
=ACC:pron-pers(M 3P ACC) se
=PIV:pp a mostrar o papel de Portugal em o mundo
CO:conj-c(<co-subj>) e
CJT:fcl
=SUBJ:np
=>N:art(M S) o
=>N:adj(<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. Essa solução é, afinal, a que segue mais estritamente o sistema CG, onde é preciso haver sempre um "núcleo" que especifica a função do nó total:

=SUBJ:np
=>N:art(M S) o
=>N:adj(<NUM-ord> M S) quinto
=P:vp é justificado

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>).

4.3 Medição do tempo de revisão

De forma a planificar o processo de construção de um ”treebank”, é necessário recolher dados sobre o tempo real de revisão. Assim, e seguindo-se o mesmo processo de revisão levado a cabo durante o projecto, um anotador reviu, no espaço de 6 horas, frases no formato CG, consagrando depois igualmente seis horas ao formato de árvores deitadas. A discussão comum de análises de casos que de alguma forma se revelaram de difícil resolução não foi efectuada.

O mesmo tempo gasto na revisão exclusivamente manual de árvores foi aplicado à revisão de árvores com o auxílio do Pica-pau (a ferramenta de auxílio de edição de árvores em emacs), ou seja três horas para cada. O objectivo não era avaliar o desempenho da ferramenta, mas apenas a medição da velocidade de revisão com o Pica-pau relativamente ao tempo gasto na revisão manual de árvores.

Os ficheiros utilizados para a medição de tempo de revisão eram idênticos em termos de número de palavras. O ficheiro revisto manualmente possuia 2685 palavras, menos 287 palavras do que no ficheiro revisto com o auxílio do Pica-pau.

Os resultados foram os seguintes: 80% do ficheiro em formato CG foi revisto.[9] Em formato de árvores deitadas, o anotador reviu manualmente (sem intervenção de ferramentas) 59% do ficheiro nas mesmas 6 horas.

Este resultado vem confirmar o facto de que a revisão das árvores revela-se de maior complexidade, uma vez que envolve níveis de revisão múltiplos. No entanto, seria de esperar que, com a revisão com o auxílio da ferramenta de edição de árvores, o processo de revisão das árvores seria mais rápido. No entanto, tal não se verificou; apenas 40% do (novo) ficheiro foi revisto.

Inspeccionaram-se então as possíveis causas de tal resultado. Um dos aspectos que, à partida, poderia ser indicador, era o número de frases em causa: com o auxílio do Pica-pau foram revistas 79 frases / árvores, enquanto que sem o seu auxílio analisaram-se 92 frases /árvores, ou seja, uma diferença de 12 frases. Apesar desta diferença o número de palavras por frase, bem como o número de constituintes por frase são bastante similares: 30,51 palavras por frase no ficheiro manualmente revisto, contra 31,14 no ficheiro revisto com auxílio do Pica-pau e 15,22 constituintes por frase contra 15,96. A complexidade das frases poderia constituir outro factor importante. Ainda que não tenha sido desenvolvida uma medida de classificação da complexidade de frases (que terá necessariamente que envolver diversos critérios distintos), as frases nos extractos revistos com o Pica-pau pareceram ser mais complexas em termos de representação no formato de árvore.

[9] Uma das frases, com 36 palavras, etiquetada incorrectamente como <s frag> e, por isso, não analisada automaticamente, foi manualmente analisada. A análise manual demorou cerca de 15 minutos, que foram tidos em conta para a contabilização total do tempo de revisão. A frase em questão não foi analisada utilizando as ferramentas de análise do VISL, porque a análise implicou a implementação de uma das opções de análise convencionadas para aquele tipo de estrutura. A frase teria de ser, de qualquer forma, revista manualmente.

Questões técnicas também poderiam ter influenciado os resultados. Nas frases revistas com o Pica-pau, é de salientar que problemas de geração de árvores a partir do formato CG, envolvendo níveis, como não indentação de nós terminais dependentes, foram mais frequentes do que nas frases submetidas a uma revisão puramente manual, fazendo com que a revisão das ditas árvores fosse mais demorada.

Finalmente, e uma vez que o desempenho da ferramenta não era o propósito deste teste, o revisor não teve em conta este aspecto, ou seja, as árvores revistas num primeiro momento com o auxílio do Pica-pau não foram verificadas posteriormente.[10] No entanto, é possível que uma operação realizada no início de uma árvore produzisse alterações indesejadas num momento posterior na mesma frase. Consequentemente, o revisor estaria, ainda que inconscientemente, a rever aspectos nas árvores derivados da utilização da ferramenta.

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).

5.1 Pica-pau, ajudante de edição

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.

O sistema é constituído por um conjunto de funções desenvolvidas em Emacs Lisp, que são incorporadas no ambiente de edição do Emacs, utilizado no projecto VISL e na Floresta. Para subir o nível (aumentar a indentação) ou baixar o nível (diminuir a indentação) de um determinado nó, o utilizador deve posicionar o cursor na linha que o contém e executar o comando apropriado. Foram criados quatro comandos de deslocamento:

a) diminui-indentação. Sobe o nível de nós terminais e de nós constituintes. No caso dos constituintes, o nó pode ser eliminado, ou seja, todos os seus descendentes passam a ser directamente dominados pelo seu sucessor directo.

b) aumenta-indentação. Baixa o nível de nós terminais e de nós constituintes. O deslocamento pode ser feito com a criação de um novo constituinte, bem como, com o deslocamento do nó para uma ramificação à esquerda ou para uma ramificação à direita

[10] Note-se, aliás, que não foi feito nenhum controlo à qualidade do resultado da revisão, nem no caso puramente manual nem no caso do uso do Pica-pau.

(em outras palavras, se o nó seleccionado deve ficar dependente do nó à sua direita ou à sua esquerda).

c) diminui-indentação-seleccionado. Diminui a indentação, **se possível**, de todos as linhas seleccionadas pelo utilizador. Este comando não verifica se estão sendo deslocados nós terminais ou constituintes. Apenas retira do início de cada linha seleccionada um sinal de igual, se houver.

d) aumenta-indentação-seleccionado. Aumenta a indentação de todos as linhas seleccionadas pelo utilizador. Este comando acrescenta ao início de cada linha seleccionada um sinal de igual, ou seja, também não verifica o tipo de nós seleccionados.

Como exemplo, para subir o nível de "=ADVL:adv('ainda') ainda", onde "ADVL:adv" é dependente da oração infinitiva cujo predicado é "fazer". A análise seria:

```
A1
STA:fcl
P:v-fin('conseguir'  PR 1S IND)    Consigo
ACC:icl
=ADVL:adv('ainda')      ainda
=P:v-inf('fazer')       fazer
=ACC:np
==>N:art('o'  M S)          o
==H:n('pino' M S)       pino
.
```

A revisão leva a trocar a posição de “=ADVL:adv“ pela de “ACC:icl”, subindo um nível. Deve-se, então, posicionar o cursor em “=ADVL:adv“ e executar o comando diminui-indentação, obtendo-se:

```
A1
STA:fcl
P:v-fin('conseguir'  PR 1S IND)    Consigo
ADVL:adv('ainda')      ainda
ACC:icl
=P:v-inf('fazer')       fazer
=ACC:np
==>N:art('o'  M S)          o
==H:n('pino' M S)       pino
.
```

Durante o primeiro teste do protótipo pela equipa de revisores, verificou-se que algumas tarefas eram constantemente executadas e que, para se chegar ao resultado desejado, o utilizador deveria interagir 2 ou 3 vezes com a ferramenta (respondendo a perguntas tais como: “insere um novo nó?” seguida de “qual o nome do nó?” e “o nó deve ser deslocado para a direita ou para a esquerda?”, se a resposta fora positiva). Então, foram criadas três especializações:

e) baixa-nível-com-inserção. Este comando tem a mesma funcionalidade do comando aumenta-indentação, mas adiciona automaticamente um novo nó (apenas o nome do novo nó é perguntado ao utilizador).

f) subir-nível-com-eliminação Este comando tem a mesma funcionalidade do comando diminui-indentação, mas elimina automaticamente o nó selecionado. Logo, funciona apenas com nós não terminais - se o utilizador tentar executá-lo sobre uma palavra ou sinal de pontuação, é exibida uma mensagem de erro.

g) subir-nível-sem-eliminação Este comando tem a mesma funcionalidade do comando diminui-indentação, mas não executa a eliminação do nó selecionado.

Para mais informações sobre cada comando, bem como um exemplo detalhado de utilização da ferramenta, veja-se Haber (2001).

5.2 Águia, sistema de procura em árvores

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.

Por isso, o desenvolvimento desta ferramenta concentrou-se na definição de sintagmas (SN, SP, etc.), na codificação das noções de dominância imediata ou (só) dominância versus precedência linear e na possibilidade de exprimir conceitos como núcleo ("head"), anteposição e posposição dentro de um sintagma.

Ainda em desenvolvimento, pretendemos expandir o Águia para a) tratar correctamente constituintes descontínuos, b) procurar (meta)anotações (tais como frases ambíguas que requerem conhecimento do mundo para a sua desambiguação) e c) apresentar um resumo quantitativo dos resultados da procura.

6 Resultado

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[11], passando pela especificação formal do formato das frases intelectualmente revistas[12] 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. Além disso, é fundamental para os que desejam prosseguir com o trabalho de expansão do ”treebank”, que o projecto esteja bem documentado em todas as suas fases.

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).

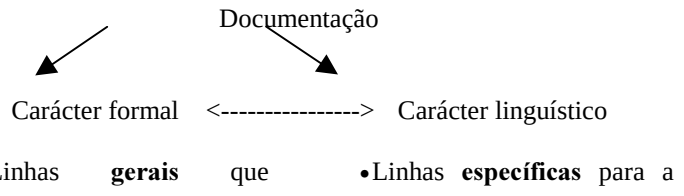
6.1 Organização da documentação

A documentação 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. Constituiu-se, para as diversas áreas da linguística, uma documentação que, durante o período de revisão, tornou mais consistente a revisão humana, e, num momento posterior, funcionará como leitura da própria Floresta Sintá(c)tica.

Apesar do carácter destas opções linguísticas ter uma componente fortemente humana (não automática), tentou-se encontrar um equilíbrio entre alterações introduzidas manual e semi-automaticamente. Ou seja, sempre que possível, as opções tomadas foram automatizadas sem comprometer a análise mais satisfatória.

A organização encontra-se sistematizada no diagrama 1:

Diagrama 1: Estrutura da documentação



[11] Presente no Glossário: http://cgi.portugues.mct.pt/treebank/glossario.html
[12] Em http://cgi.portugues.mct.pt/treebank/BNFfloresta.html.

presidem à anotação; • Não restrito ao projecto da Floresta Sintá(c)tica; princípios que regem o projecto VISL.	revisão da análise automática; • Tomada de decisão quanto ao procedimento a ser seguido no projecto nos diversos níveis de análise.

6.2 Processo de documentação das opções linguísticas

O processo de documentação compreende quatro fases distintas: a primeira frase refere-se à identificação de casos que sejam problemáticos em termos de análise. Esses casos compreendem diferentes questões: tópicos não ”convencionais” do português, não descritos nas gramáticas tradicionais; tópicos que envolvem mais do que uma dimensão linguística (tópicos pragmático-discursivos, por exemplo); problemas derivados do uso incorrecto / impreciso da pontuação; complexidade na representação formal devido à própria estrutura frásica; presença de listas (enumerações).

Após a identificação dos casos-problema por cada anotador, passa-se à discussão em comum com a equipa encarregada da parte de revisão linguística, com o objectivo de encontrar uma ou mais análises para cada um dos casos. No entanto, esta decisão deve ser sistematizada / generalizada, ou seja, não faz sentido encontrar soluções para casos isolados. É, por isso, fundamental que se proceda a uma busca no corpus de forma a confirmar a aplicabilidade da solução.

O passo seguinte é a materialização deste processo em suporte escrito, ou seja, a própria documentação, base para a aplicação sistemática das opções acordadas.

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 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 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 recorrência à sua forma gráfica, por razões práticas, uma vez que era o mesmo conjunto de frases que estariam a ser visualizadas 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 conclusões, consulte-se Afonso (2001).

Visto que o teste cobria toda a revisão, muitos parâmetros estavam em jogo, o que tornou a sua análise numa tarefa muito longa e a interpretação dos resultados difícil.

Os resultados em sintaxe, forma e função, apresentam-se na tabela I, em que as percentagens são referentes ao **número total de diferenças** contabilizadas, não às 107 frases revistas:

Tabela I: Contabilização das diferenças observadas em sintaxe

		Forma: n° diferenças p/ frase				Função: n° diferenças p/frase
		Indentação	Nó extra / ausência nó obrigatório	Posição nó	forma	
		133 51	116 41	27 15	84 32	303 74
RAZÕES DAS DIFERENÇAS	Erro humano	95,4%	80,4%	62,9%	82,1%	77,8%
	≠ análises aceites	3%	6,9%	37%	17,8%	15,8 %
	≠ análises não aceites	–	–	–	–	0,66%
	Outras análises (uma em comum)	1,5%	0,9%	–	–	5,61%
	Variante do port.	–	–	–	–	–
	Conhecimento do mundo (extra-linguístico)	–	–	–	–	–

Como se pode ver pela tabela I, de um total de 360 diferenças observadas em forma sintáctica, a maioria situa-se a nível da indentação (133, uma média de 2,6 diferenças por frase) e da ausência de nó obrigatório / nó extra (116 diferenças, uma média de 2,8 diferenças por frase). Em termos de função sintáctica, observaram-se 303 diferenças distribuídas por 74 frases (uma média de 4 diferenças por frase). Em todas as categorias tanto de forma como de função sintáctica, a causa predominante das

diferenças observadas foi erro humano, categoria que abrange quer incorrecções na atribuição de uma dada classificação/função, quer falta de atenção.

Em morfologia, os resultados encontram-se na tabela II abaixo:

Tabela II: Contabilização das diferenças observadas em morfologia

		Classe de palavras (n° diferenças p/ frase)	Lema (n°diferenças p/ frase)	Género e número (n° diferenças p/ frase)
		15 / 12	3 / 2	74 / 72
Razões das diferenças	Erro humano	100%	33,3%	94,5%
	≠ análises aceites	-	-	-
	≠ análises não aceites	-	-	-
	Outras análises (uma em comum)	-	-	-
	Variante do port.	-	66,6%	-
	Conhecimento do mundo (extra-linguístico)	-	-	5,4%

Pelos dados da tabela II, conclui-se que a categoria onde se registou o mais elevado número de diferenças foi a classe de palavras (15 diferenças em 12 frases) por erro humano, o que não deixa de ser um facto curioso, pois à partida não haveria razão para haver dúvidas sobre essa matéria.

Investigaram-se, por isso as razões que poderiam explicar a elevada percentagem de erro humano e de entre as causas prováveis estão a complexidade e tamanho das frases, a alteração no processo de revisão: a não visualização das árvores invertidas (gráficas) e a revisão directa em formato de árvores. Ou seja, é provável que o facto de o anotador rever toda a informação (morfossintáctica e estrutural – níveis de contituintes) no formato de árvores deitadas tenha contribuído para o aumento do número de diferenças observadas nas diversas categorias. Este facto é corroborado, em parte, pela percentagem de erro humano no que respeita às diferenças na categoria género e número. Ou seja, estes valores significam falta de desambiguação por parte dos anotadores, o que pode indicar que descuraram a revisão desta parte da morfologia em detrimento da revisão da sintaxe.

Além disso, a revisão das árvores com a ferramenta desenvolvida, mas não completamente operacional, poderia ter produzido alterações não desejadas nas árvores.

Mas mais do que os resultados quantitativos do teste, foram importantes as reflexões que daí surgiram sobre futuros desenhos de testes de consistência, bem como sobre todo o processo de construção de um ”treebank”.

A primeira conclusão relaciona-se com o objecto de medição. Talvez tivesse sido mais produtivo restringir o objecto de medição a poucos parâmetros, essencialmente aqueles que se suspeitaria que à partida levassem a um maior número de diferenças. Segundo Brants:

> Categories which cause large number of differences are good candidates for improving inter-annotator agreement. A better handling of these categories has the potential of eliminating large number of differences (Brants, 2000)

Além disso, segundo o mesmo autor há questões de mais fácil avaliação do que outras.

Acrescentar-se-ia ainda que seria aconselhável a realização faseada deste tipo de testes de avaliação ao longo de todo o processo de revisão.

Relativamente à contagem, e de forma a controlar os resultados da contagem seria talvez útil realizar estudos prévios sobre o tipo de resultados possível, como, por exemplo, quantas diferenças máximas e / ou mínimas poderiam ser obtidas na comparação.

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. Deste trabalho de exploração emergiram algumas ideias de como conceber um ”treebank”, tendo em conta os mais diversos factores como o(s) objectivo(s) do próprio projecto, o tamanho e tipo / características do corpus, recursos humanos disponíveis, factor tempo.

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 constituiram a base de todo o trabalho de recolha de tipos de problemas, bem como a 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 possibilidade de análise, critérios de ”plantação” de árvores, resolução de casos problema, prepara-se uma futura revisão do corpus de forma mais sólida / consistente. Além disso, fruto de este trabalho inicial, o analisador automático foi-se adaptando também às novas necessidades, sendo progressivamente melhorado.

Por isso, parece-nos que uma proposta válida para a revisão de um corpus anotado maior seria a sua divisão em partes com diferentes níveis de especificação e também de perfeição (percentagem de erros), ou seja, uma das partes, por exemplo 10%, seria sujeita a uma revisão exaustiva, a todos os níveis; outra(s) parte(s) (até 50%) seria(m) revista(s) tendo em conta determinadas categorias a nível da oração principal (Sujeito, Complemento directo, Predicativo do sujeito/objecto, etc.). A análise do resto do corpus seria exclusivamente automática, isto é, não revista. No entanto, a análise automática contaria, nesta altura, com os melhoramentos derivados de uma revisão exaustiva nos 10%, que teriam implicado, como já referido, uma discussão de casos-

problema, amplificação do léxico, adaptação do analisador sintáctico relativamente ao corpus em questão.

No entanto, é fundamental ter em consideração um compromisso entre qualidade/ambição e quantidade/realidade. Em suma, o projecto da Floresta Sintá(c)tica foi talvez muito ambicioso para um primeiro ano de existência. Optou-se pela qualidade, inserindo-se distinções mais finas, menos automáticas e mais humanas e, consequentemente, menos sistemáticas, além de, em alguns casos, colidirem com categorias já existentes, o que produziu um maior grau de erro.

O grau de erro inter e intra-anotadores deve ser diminuído (idealmente eliminado), sendo o grau de consistência uma das prioridades de um projecto deste tipo, porque a consistência na anotação aumenta consideravelmente a utilidade de um corpus tanto para o teste de analisadores morfossintácticos, como para a investigação linguística (Brants, 2000).

Uma das formas de reduzir o número de diferenças por ”erro humano” seria o desenvolvimento de ferramentas de auxílio à revisão que impedissem, marcassem para inspecção posterior ou corrigissem certas operações porque incompatíveis em termos formais. O treino de anotadores numa fase de pré-construção do ”treebank” também poderia incrementar níveis de consistência, bem como a realização de testes de consistência periódicos e restritos a poucos parâmetros de cada vez. Desses testes, retirar-se-iam situações recorrentes para cuja colmatação se estudaria depois a melhor estratégia.

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.

Brants, Thorsten. "Inter-Annotator Agreement for a German Newspaper Corpus", in Maria Gavriladou, George Carayannis, Stella Markantonatou, Stelios Piperidis & Gregory Stainhaouer (eds.), *Proceedings of the Second International*

Conference on Language Resources and Evaluation, LREC2000 (Athens, 31 May-2 June 2000), pp. 1435-1439, Vol. 3.

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", *Third Workshop on Linguistically Interpreted Corpora*, http://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), pp.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.

Haji•, 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, pp. 313-30.

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

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), pp.131-140.

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), pp.105-113.

Santos, Diana & Eckhard Bick. "Providing Internet access to Portuguese corpora: the AC/DC project", in Maria Gavriladou, George Carayannis, Stella Markantonatou, Stelios Piperidis & Gregory Stainhaouer (eds.), *Proceedings of the Second International Conference on Language Resources and Evaluation, LREC2000* (Athens, 31 May-2 June 2000), pp.205-210.

APÊNDICES

A. O analisador automático PALAVRAS

PALAVRAS é um analisador automático (tagger-parser) para português escrito ou transcrito, e algumas aplicações nele baseadas, (a) uma ferramenta para tratamento de corpora e (b) uma ajuda para ensino de sintaxe através da internet. O sistema apoia-se num léxico com 50.000 lemas e milhares de regras gramaticais para fornecer uma análise completa, tanto morfológica como sintática, de um texto qualquer.

Para otimizar robusteza, eficiência e recall, as ferramentas gramaticais internas do sistema foram baseadas no formalismo Constraint Grammar (Constraint Grammar, CG), introduzido por Fred Karlsson (1990, 1995). O núcleo original de programas foi construido em cima dum parser multi-nível desenvolvido no contexto dum projeto de doutoramento na universidade de Århus (Bick 1996, 1997). Embora usando um conjunto de etiquetas (tag set) bastante amplo, o parser alcança – com textos desconhecidos - um nível de correteza (correctness) de 99% em termos de morfologia (classe de palavras e flexão), e 97-98% em termos de sintaxe.

Entre outras, dependência, estrutura de argumentos e orações subordinadas são tratadas de uma maneira inovadora, permitindo a transformação automática da notação sintática chata de dependência de CG em árvores sintáticas de constituintes, mais atrativas visual e pedagogicamente. O parser utilisa informação lexical sobre regência e classe semântica, e estudos preliminares sugerem a possibilidade de desambiguação nesse nível também.

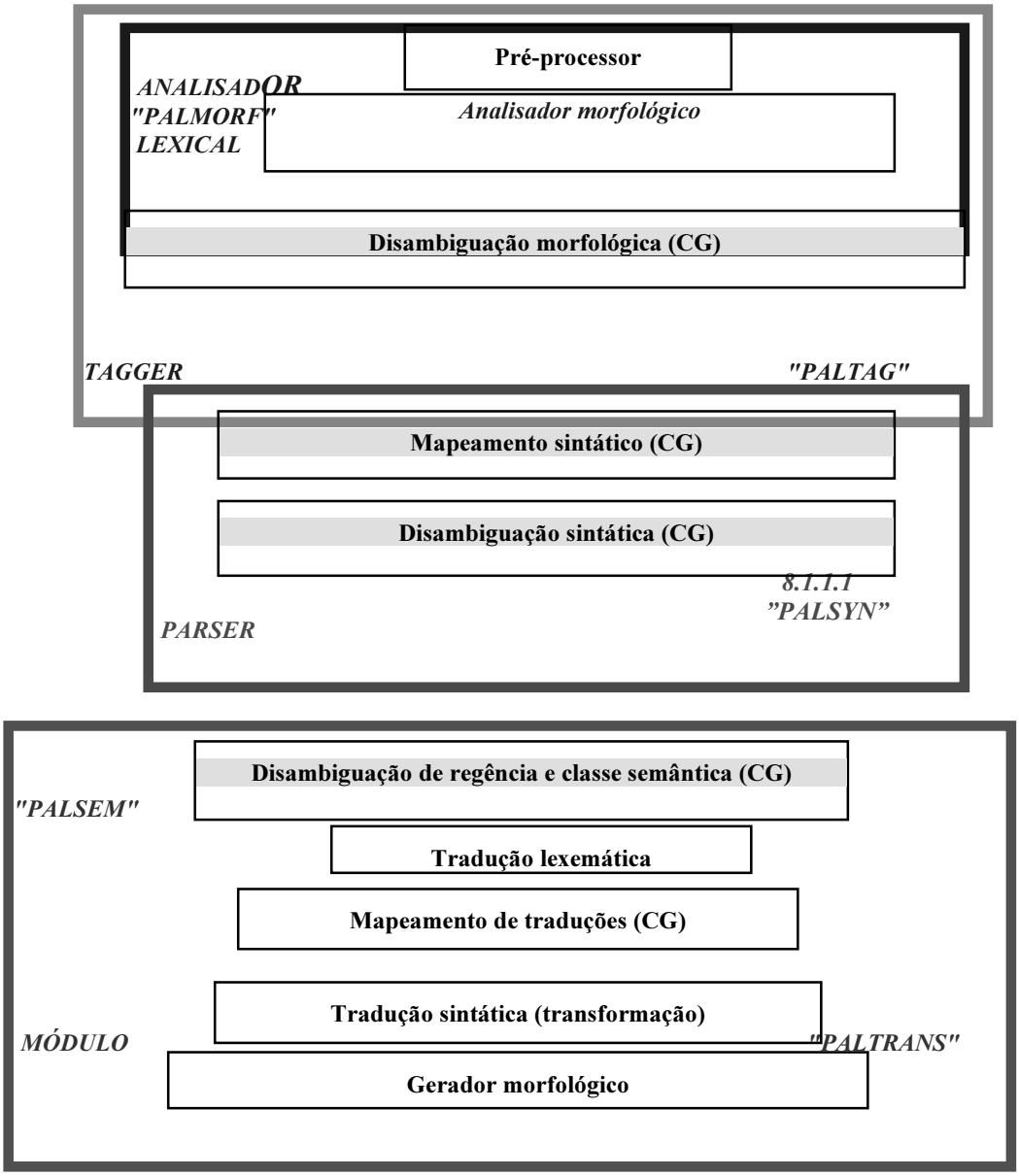
Na interface de ensino gramatical, os usuários podem escolher entre vários filtros notacionais, apoiando diferentes paradigmas descritivas da língua. Exemplos são exercícios nos quais coram-se palavras para marcar sua classe, ou árvores de sintaxe gráficas construídas e manipuladas pelo estudante e controlado pelo computador, com etiquetas de forma e função em cada nó.

Quando usado para etiquetagem de corpora "crus" (não anotados), o parser permite buscas complexas, juntando ao mesmo tempo palavras e lemas, classe de palavra e função sintática. Fora de aplicações óbvias como lexicografia e linguística, a ferramenta de corpus pode integrar-se na interface de ensino, fornecendo ao estudante exemplos de certas estruturas gramaticais.

O sistema analisa ca. 200 palavras/sec num computador Pentium/Linux quando trabalhando com todos os níveis. Morfologia, só, alcança 2000 palavras/sec.

Estrutura modular e hierárquica do parser

Tecnicamente, o parser funciona como um conjunto hierárquico de módulos morfológicos, sintáticos e semânticos, tão lexicais como disambiguacionais, cada um tirando e ameliorando output do módulo anterior e fornecendo input para o módulo posterior. Num contexto aplicativo de tradução automática, por exemplo, usam-se os seguintes níveis:


Pré-processor
ANALISADOR
"PALMORF"
LEXICAL
Analisador morfológico
Disambiguação morfológica (CG)
TAGGER
"PALTAG"
Mapeamento sintático (CG)
Disambiguação sintática (CG)
8.1.1.1
"PALSYN"
PARSER
Disambiguação de regência e classe semântica (CG)
"PALSEM"
Tradução lexemática
Mapeamento de traduções (CG)
Tradução sintática (transformação)
MÓDULO
"PALTRANS"
Gerador morfológico

Corpus CETEMPúblico

O CETEMPúblico (Corpus de Extractos de Textos Electrónicos MCT/Público) é um corpus de aproximadamente 180 milhões de palavras em português europeu, criado pelo projecto Processamento computacional do português após a assinatura de um protocolo entre o Ministério da Ciência e da Tecnologia português (MCT) e o jornal PÚBLICO em Abril de 2000. O corpus inclui o texto de cerca de 2.600 edições do PÚBLICO, entre os anos de 1991 e 1998, num total de aproximadamente 180 milhões de palavras (versão 1.0).

O CETEMPúblico 1.0 está dividido em 1.567.625 extractos, classificados por semestre e secção do jornal da qual provêm. Cada extracto está dividido em parágrafos e frases, e os títulos e os autores dos artigos estão assinalados.

Considerámos palavras todos os tokens existentes no corpus que contenham pelo menos uma letra ou dígito. Os sinais de pontuação não foram contados como palavras. A categoria "Pontuação" inclui as unidades com sinais de pontuação, tal como *(1993)*, *a)* ou *17:53*.

Dados quantitativos aproximados referentes à versão 1.7 encontram-se aqui:

	Formas	Tipos
Unidades	229.038.019	1.033.041
Palavras	191.687.833	999.059
Pontuação	13.065.151	33.982

Quanto à estrutura mais fina do corpus, veja-se a seguinte tabela;

Estrutura	Número
Extractos <ext>	1.504.258
Parágrafos <p>	2.571.735
Frases <s>	7.082.094
Títulos <t>	655.059
Autores <a>	247.392
Elementos de lista	80.060

Este corpus destina-se primariamente a todos quantos desenvolvem programas que processam a língua portuguesa, e que consequentemente precisam de matéria prima para o seu trabalho. A versão em formato texto distribuída em CD destina-se principalmente a este tipo de investigadores.

Por outro lado, espera-se que o corpus seja útil a todos os estudiosos da língua que queiram confirmar as suas hipóteses em material textual, previamente organizado. As versões CQP e acesso através da rede foram pensadas para este público alvo que, no entanto e se assim o desejar, também pode obter o CD para o ter localmente, e eventualmente codificar o corpus no sistema de processamento de corpora da sua preferência.