Calidoscópio
Vol. 6, n. 3, p. 142-148, set/dez 2008

Claudia Freitas
maclaudia.freitas@gmail.com

Paulo Rocha
paulo.rocha@di.uminho.pt

Eckhard Bick
eckhard.bick@mail.dk

Um mundo novo na Floresta Sintá(c)tica – o treebank do Português

A new world in Floresta Sintá(c)tica – the Portuguese treebank

RESUMO – A Floresta Sintá(c)tica tem como objetivo criar e disponibilizar um *corpus* sintaticamente anotado. Neste artigo, são apresentados dois novos materiais do projeto: *Selva* (300 mil palavras e parcialmente revisto) e *Amazônia* (3.8 milhões de palavras, não revisto). Para lidar com um material tão grande e variado foi construída a interface Milhafre. O artigo mostra, ainda, como vem sendo enfrentado o desafio de compatibilizar, de uma lado, o usuário lingüista, que pode ter um perfil muito heterogêneo e, em geral, pouca familiaridade determinadas formalizações mais utilizadas em informática e, de outro, um único modelo de anotação sintática, freqüentemente pouco conhecido do lado "lingüístico não-computacional" e uma interface de acesso e manipulação de corpora capaz de lidar com um objeto tão complexo como a língua.

Palavras-chave: árvores sintáticas, *corpus* anotado, *corpus* revisto, busca em corpora.

ABSTRACT – *Floresta Sintá(c)tica* is a publicly available treebank for Portuguese, created as a collaboration project between *Linguateca* and the VISL project. It consists of Brazilian and European Portuguese texts automatically annotated by the parser PALAVRAS (Bick, 2000) and manually revised. In this paper, we present two new corpora, *Selva* (composed by literary, scientific and transcribed spoken texts, partially revised) and *Amazonia*, (a huge *corpus* of 3.8 million words, unrevised), and a user-friendly web based *corpus* tool, Milhafre. We also present how we manage to balance (a) our user, which can have different linguistic background, (b) the need for a grammar that is rich and complex enough in order to process real language (our corpora); and (c) the absence of a consensual syntactic model.

Key words: Portuguese treebank, annotated *corpus*, revised *corpus*, user-friendly *corpus* tool.

Introdução

Uma floresta sintática – tradução do inglês *treebank* – é um conjunto de itens (frases) analisados sintaticamente. A cada frase é atribuída uma estrutura sintática hierárquica, e por isso uma frase (sintaticamente analisada) pode ser vista como uma árvore, donde uma floresta nada mais é que um conjunto de frases analisadas sintaticamente e com informação relativa aos níveis de constituintes. Florestas sintáticas costumam ser utilizadas, de maneira geral, tanto em estudos da língua baseados em *corpus* como no treino de analisadores sintáticos.

O projeto Floresta Sintá(c)tica é uma iniciativa dos projetos Linguateca[1] e VISL[2] (Visual Interactive Syntax Learning) que tem como objetivo a criação, revisão lingüística e disponibilização de um *corpus* da língua portuguesa (do Brasil e de Portugal) analisado sintaticamente. O projecto, iniciado em 2000, vem sendo ativamente retomado desde 2007[3]: a partir de então, foi acrescentado novo material (novos textos) e novas etiquetas sintáticas, uma revisão ainda mais fina teve lugar e foi desenvolvida uma nova interface de busca em árvores sintáticas, o Milhafre.

O objetivo deste artigo é não apenas (re)apresentar a Floresta Sintá(c)tica como um recurso valioso para aque-

[1] Disponível em http://www.linguateca.pt.
[2] Disponível em http://visl.sdu.dk/.
[3] Mais informações sobre a história do projeto Floresta Sintá(c)tica encontram-se em http://www.linguateca.pt/Diana/download/SantosBickAfonsoFlorestaSet2007.pdf3

les que trabalham com a língua portuguesa, mas fazê-lo sob duas perspectivas distintas: em um primeiro momento, descrevemos o processo de criação do *corpus* da Floresta e exemplificamos brevemente alguns de seus usos, para em seguida explicitar as idéias que norteiam o projeto e como isso se manifesta na prática; em um segundo momento, descrevemos a interface de busca Milhafre, criada para facilitar a pesquisa em árvores sintáticas.

Floresta Sintá(c)tica: como e por quê

Como é feita a Floresta Sintá(c)tica

Florestas sintáticas podem ser criadas de uma maneira completamente automática, ou semi-automática, com uma primeira análise (ou anotação) automática, seguida de revisão. Na Floresta, temos os dois tipos de abordagem: material bruto, isto é, sobre o qual atuou apenas o analisador sintático PALAVRAS (Bick, 2000); material integralmente revisto por lingüistas, o *Bosque*; e material apenas parcialmente revisto por lingüistas, a *Selva*.

De maneira geral, o processo de criação da Floresta segue os seguintes passos: o texto "cru", isto é, sem qualquer tipo de anotação, é enviado para o analisador sintático PALAVRAS, que devolve o texto já anotado. É sobre esse material - um arquivo texto, chamado versão AD (árvores deitadas) – que se dá a revisão lingüística. A Figura 1 ilustra um pedaço de árvore (saída do PALAVRAS), em que se pode visualizar a hierarquia entre as unidades sintáticas.

```
Seca afeta pouco a produção de grãos

STA:fcl
=SUBJ:np
==H:n('seca' <np-idf> F S)  Seca
=P:vp
==MV:v-fin('afetar' PR 3S IND)      afeta
=ADVL:advp
==H:adv('pouco' <quant>)  pouco
=ACC:np
==>N:art('o' <artd> F S)    a
==H:n('produção' F S)       produção
==N<ARGO:pp
===H:prp('de')     de
===P<:np
====H:n('grão' M P)         grãos
```

Figura 1. Extrato de árvore no formato AD.

No formato AD, a informação lingüística é codificada por meio dos pares função e forma. Assim, na segunda linha da Figura 1, há a indicação de que à função sujeito (SUBJ) corresponde a forma sintagma nominal (np). A informação relativa à hierarquia é marcada com o sinal de =. Na figura, vemos que o sujeito "seca" está no mesmo nível (e, portanto, é um argumento de) do predicado (P) "afeta" e também está no mesmo nível do adjunto adverbial ("pouco") e do objeto direto (ACC) " a produção de grãos". O núcleo deste objeto direto (H:n), "produção", contém ainda, por sua vez, o complementos ("de grãos").

Nem todas as etiquetas sintáticas usadas pelo PALAVRAS são utilizadas na Floresta. Por isso, é preciso uma etapa intermediária de pareamento entre etiquetas. Além desse formato AD, há material da Floresta - nomeadamente o *Bosque* e a *Floresta Virgem* - disponível em outros formatos, como SQL, Penn Treebank, Perl, XML e Tiger XML.

A utilidade de uma floresta sintática

Embora a utilidade de uma floresta sintática - e, especificamente, da Floresta Sintá(c)tica - já tenha sido apresentada em Afonso (2004), listamos novamente aqui algumas das possibilidades: descrição da língua, ensino, treino de analisadores morfossintáticos e avaliação do desempenho de sistemas em anotação morfológica de corpora.

Em uma perpectiva voltada para o ensino de língua, o uso de corpora – exemplos de textos reais – em sala de aula pode ser um importante aliado na medida em que, ao colocar o aluno em contato com uma ampla variedade de fatos da língua, incentiva a reflexão e o questionamento sobre a língua, colocando-o como sujeito de investigação, capaz de depreender regularidades, irregularidades e, em última instância, capaz de perceber a língua como um fenômeno multifacetado. Além disso, a possibilidade de confrontar "informação" teórica – proveniente de gramáticas, por exemplo – com porções de texto reais é um estímulo ao espírito crítico e à percepção da língua como uma entidade viva. Neste ponto, o uso não apenas de corpora, mas especialmente de corpora *anotados* e lingüisticamente *revistos* faz da Floresta um recurso com grande potencial a ser explorado. Em termos concretos, com o auxílio de uma interface de busca em árvores (como será visto mais à frente), é possível, por exemplo, encontrar orações em que não há uma concordância gramatical entre sujeito e predicado, e confrontar tais frases com o que dizem as gramáticas; encontrar orações passivas em que o agente da passiva é omitido, a fim de refletir sobre o efeito no texto de tais construções; investigar transitividade de nomes e de verbos e o uso de tempos verbais, dentre uma série de outras possibilidades.

Em uma perspectiva informática, o *corpus* da Floresta pode ser de grande valia para o treino de analisadores morfossintáticos: há um vasto material anotado (e agora ainda mais, com a criação da *Selva* e da *Amazônia*), há mais informação lingüística disponível) e é possível coletar material específico para o treino de determinadas

estruturas. Isto é, assumindo que, em aprendizagem automática, nem sempre mais significa melhor (Ji e Grishman, 2006), é bastante simples conseguir uma lista de frases em que ocorrem determinados fenômenos. Por exemplo, a ambigüidade relativa ao local de encaixe do sintagma preposicionado costuma ser um ponto fraco em diversos sistemas, justamente porque é de uma estrutura potencialmente ambígua na língua. Nesse caso uma busca apenas por frases que contêm SPreps atrelados a nomes (ou a verbos, ou a ambos), facilmente realizada com a interface Milhafre, pode viabilizar a criação de um material de treino mais eficaz.

A filosofia florestal

Porém, para que a Floresta possa, efetivamente, servir aos objetivos pretendidos, é importante que siga o que chamaremos aqui de "filosofia florestal", isto é, diretrizes que norteiam as escolhas lingüísticas da Floresta. Esta filosofia surge da intenção de *refletir um consenso entre as possibilidades de análise sintáctica de um fenômeno*, ou pelo menos, permitir uma escolha informada (Afonso *et al.*, 2001). Em consequência, são objetivos da Floresta (i) oferecer material para uma vasta gama de usuários, de lingüistas a engenheiros; (ii) servir de espaço para a investigação, e não para demonstrações de correntes teóricas específicas - embora, obviamente, não se possa fugir a teorias subjacentes à anotação.

A filosofia florestal é, portanto, uma tentativa de equilibrar (i) a necessidade de uma gramática rica e complexa o suficiente para dar conta (leia-se analisar automaticamente) dos exemplos reais da língua; (ii) a ausência de unanimidade com relação a modelos sintáticos e (iii) a formação lingüística dos utilizadores (que deve poder ser mínima). Por isso, embora subjacente ao PALAVRAS esteja o modelo " Constraint Gramar", o que procuramos oferecer ao utilizador – lingüista, engenheiro ou interessado –, sempre que possível, é informação que, do ponto de vista da terminologia, tenha pontos em comum com a nomenclatura usada pela gramática tradicional[4]. Embora limitações e inconsistências da GT já tenham sido amplamente discutidas (Perini, 1986; Franchi, 2006), trata-se da tradição mais difundida, o que possibilita maior autonomia e facilidade de uso: para os que têm pouca familiaridade com termos como *objeto direto*, ou *agente da passiva*, uma consulta às gramáticas é algo simples; para os que trabalham com outros modelos sintáticos, é possível perceber que tipo de informação está codificado na Floresta, e transpô-lo para o modelo desejado. Para tanto, consideramos que uma documentação consistente de todas as opções lingüísticas tomadas durante a anotação é de fundamental importância, para que se possa, como já dissemos, permitir uma escolha informada[5].

Além da preocupação com a nomenclatura, e ainda mais importante que isso, a filosofia florestal se manifesta na busca, sempre que possível, por uma descrição dos fenômenos lingüísticos a partir de um alto nível de granularidade[6], pois é mais fácil agrupar fenômenos que podem, eventualmente, ser tratados de maneira disjunta, do que manualmente distinguir casos.

A seguir exemplificamos como a filosofia florestal norteia as escolhas lingüísticas da Floresta Sintá(c)tica.

A filosofia na prática: alguns exemplos de anotação

Argumentos versus modificadores do nome

Até há algum pouco tempo , argumentos e modificadores do nome recebiam o mesmo tratamento na Floresta: todos eram igualmente considerados modificadores de nome (e recebiam a etiqueta N< ou >N, consoante a direção do núcleo nominal que estivesse a modificar, isto é, "prefeito corrupto" e "novo prefeito", respectivamente).

Porém, tendo em vista a língua portuguesa, consideramos interessante uma distinção mais fina dessa categoria, o que corresponde à distinção, em temos tradicionais, entre adjunto adnominal (modificadores do nome) e complemento nominal (argumentos do nome). Tal distinção pode ser relevante não apenas de uma perspectiva de descrição lingüística, mas também na identificação da estrutura argumental de nomes deverbais, que por sua vez permite a identificação de papéis semânticos (e pode auxiliar a análise semântica de textos).

A distinção entre complementos nominais (CN) e adjuntos adnominais (AA), contudo, não é unânime (para uma discussão aprofundada do tema, sugerimos a leitura de Azeredo, 2001 e Meyer, 1995). De maneira simplificada, temos que complementos nominais correspondem aos objetos de nomes deverbais, ou seja, se temos "comprar a casa", em "a compra da casa", o "da casa" é um complemento do nome. Já em "a chegada do inverno", o "do inverno" não seria complemento, mas adjunto, porque corresponde ao sujeito do verbo intransitivo "chegar". O problema é que, além da definição não ser consensual,

[4] Falamos aqui em "gramática tradicional" em termos genéricos, pois sabemos que também não há "uma" gramática tradicional.

[5] Chamamos *Bíblia Florestal* (Freitas e Afonso, 2008) ao documento que contém não só as opções lingüísticas da Floresta, como também um glossário com todas as etiquetas utilizadas.

[6] A referida granularidade também depende das possibilidades de análise realizadas pelo parser PALAVRAS pois, dado o vasto *corpus*, a criação/revisão manual de todas as etiquetas seria um trabalho sem fim.

há situações em que os próprios exemplos contradizem a definição.

Ao invés de tomarmos uma decisão que seria, inevitavelmente, controversa – quer escolhêssemos considerar apenas CN os objetos de nomes deverbais, quer considerássemos CN qualquer argumento de um nome deverbal – preferimos uma solução que distingue entre 4 casos de "modificadores":

(a) argumentos do nome que ocupariam a posição de sujeito: chegada do inverno (= o inverno chegou). Este caso é marcado com a etiqueta N<ARGS[7].
(b) argumentos do nome que ocupariam a posição de objeto: compra da casa (= comprar a casa). Este caso é marcado com a etiqueta N<ARGO.
(c) argumentos do nome que não se enquadram em nenhum dos casos anteriores, pois são nomes que se relacionam a adjetivos (a possibilidade de compra) ou nomes transitivos como medo (medo de barata), e outros ainda que se enquadram em construções partitivas (monte de gente). Estes casos são marcados com a etiqueta N<ARG.
(d) modificadores do nome (ou adjuntos adnominais), como "a compra de ontem", "a entrada dos fundos", "um monte de 5.000 metros de altitude". Esses casos são marcados com a etiqueta N<.

Retomando a filosofia da Floresta, a melhor escolha de anotação é a mais produtiva, isto é, aquela que é capaz de atender ao maior número possível de usuários, desde que embasada um uma perspectiva teórica relativamente consensual. Se não assumíssemos essa posição, com quatro possibilidades distintas de anotação, deixaríamos o investigador da língua, interessado sobre (a distinção entre) essas classes ou desejoso de fundamentar uma ou outra perspectiva, com um trabalho árduo. Por exemplo, caso desejasse fazer uma comparação entre complementos do nome do "tipo sujeito" e do "tipo objeto", teria que, ao lado dos complementos do tipo objeto (os CN "tradicionais") obtidos pela busca no *corpus*, procurar ainda todos os outros modificadores de substantivos introduzidos por preposição (extremamente abundantes na língua) e, a partir daí, filtrar, manualmente, o que lhe interessasse. Da forma como a anotação foi feita, porém, basta juntar as ocorrências das quatro etiquetas.

Partitivos

O conceito de partitividade é essencialmente um conceito semântico, relacionado à habilidade humana de contagem e medição (Peres, 1992). Sendo assim, a tentativa de anotar a sua contraparte sintática encontra, inevitavelmente, diferentes critérios – aos quais correspondem diferentes perspectivas sobre os partitivos – que podem se sobrepor. Assim, sob o rótulo de partitivos, é possível distinguir diferentes fenômenos, de acordo com o que se está procurando: é possível usar a partitividade para sustentar a distinção (semântica) entre contáveis e não contáveis, para investigar diferenças relativas a uma concordância com base em um núcleo sintático *versus* um núcleo semântico ("metade dos deputados não compareceu" *versus* "metade dos deputados não compareceram"), ou ainda a idéia mais geral de parte de um todo. A fim de abarcar esses diferentes interesses, procuramos manter tais distinções, introduzindo três etiquetas diferentes para cada um desses casos.

Nomes *versus* adjetivos

Por fim, o último exemplo se refere à distinção (ou flutuação) entre a classe dos substantivos e a dos adjetivos. Em muitos casos, é notória a dificuldade de classificação, com bons argumentos tanto para os que sustentam a presença de uma "classe" dos adjetivos substantivados (Basílio, 2004), como para os que sustentam um caso "genuíno" de flutuação ou mesmo a existência de uma classe única nominal (Perini, 1997). Na Floresta, procuramos, mais uma vez, apontar para a existência de tais situações limite, mas sem tomar partido por nenhuma das opções, disponibilizando material para a investigação. Assim, em casos como os abaixo, utilizamos a etiqueta n-adj, que indica que o item tanto pode ser visto como um adjetivo ou como um substantivo, mas que não somos capazes de estabelecer a distinção:

(a) integrante da campanha;
(b) em favor dos evangélicos;
(c) os responsáveis pelos menores;
(d) abuso sexual em bulímicas.

As três situações exemplificadas (argumentos de nome, partitivos e n-adj) buscaram não apenas mostrar como tentamos levar para a prática da análise sintática a idéia de "espaço de pesquisa", mas fazê-lo por meio da apresentação de algumas novas etiquetas da Floresta.

Lembramos ainda que, embora talvez nunca formulada de maneira tão explícita, a filosofia florestal desde sempre esteve presente nas escolhas, principalmente, de revisão das árvores sintáticas. Por exemplo, no caso de haver diferentes possibilidades de análise sintática, buscou-se sempre representar todas as árvores correspondentes às análises possíveis. As etiquetas SA/OA/ADVL, referentes a adjuntos adverbiais relacionados ao sujeito

[7] A etiqueta N<ARGS significa que se trata de um ARGumento relativo ao Sujeito, e que está à direita do núcleo nominal N<.

("a Rússia está *perto da estabilidade política*"), objeto ("o governo mandou uma alteração *ao congresso*") e livres ("isto aconteceu *ontem*"), respectivamente, buscam também refinar a possibilidade de descrição dos adjuntos adverbiais, aproveitando classificações mais finas já produzidas pelo *parser* PALAVRAS.

Por fim, acreditamos que o trabalho do lingüista, durante o processo de anotação/revisão de *corpus* no âmbito de um projeto como a Floresta Sintá(c)tica, que se propõe a servir a comunidade não apenas de PLN, mas também lingüística, é, acima de tudo – de maneira ideal – um trabalho de busca da invisibilidade. Isto é, no contexto do projeto, mais importante do que demonstrar a viabilidade ou sucesso de um determinado um modelo de língua, é fazer, do espaço, um local em que diversos modelos possam compartilhar igualmente a chance de serem ou não corroborados. E tentar ser invisível é isso: é não impor uma visão (o que sem dúvida é tentador para quem anota/revê), mas deixar que o trabalho de outros sobre o *corpus* aconteça da maneira mais natural possível. Sabemos, é claro, que isso é impossível, como toda e qualquer outra tentativa de neutralidade, mas ainda assim é um norte do qual procuramos não nos afastar muito.

De que é feita a Floresta Sintá(c)tica? As partes da Floresta

O projecto Floresta Sintá(c)tica se subdivide, atualmente, em 4 partes, que descrevemos a seguir[8]:

Floresta Virgem: Contém cerca de 1.6 milhões de palavras (95 mil frases) coletadas do início dos corpora CETENFolha (parte do *corpus* NILC/São Carlos, retirado de textos do jornal brasileiro *Folha de São Paulo*, de 1994) e CETEMPúblico (retirados do jornal português PÚBLICO) e anotadas automaticamente. Não contém qualquer tipo de revisão linguística. (A Floresta Virgem não contém as frases pertencentes ao *Bosque*).

Amazônia: Contém 3.8 milhões de palavras (cerca de 194 mil frases) retiradas do sítio colaborativo Overmundo, um coletivo virtual que tem como objetivo expressar a produção cultural brasileira. Por ser colaborativo, o sítio conta com um grande número de autores, de diversos pontos do Brasil, o que se reflete também em diferentes estilos de escrita. Para a Amazônia, foram coletados textos da seção "Overblog" e textos de não-ficção da seção "Banco de Cultura" disponíveis em 30 de Setembro de 2008, perfazendo um total de 4070 textos (e 1303 autores). Diferentemente dos outros corpora da Floresta, a Amazônia não é um *corpus* balanceado entre o português do Brasil e de Portugal: todos os textos são brasileiros. E, assim como a Floresta Virgem, a Amazônia também é um *corpus* sintático em estado bruto – a sua anotação não foi revista por lingüistas.

Selva: Contém cerca 300 mil palavras divididas entre diferentes modalidades (escrita e falada), gêneros, domínios e as variantes portuguesa e brasileira do Português. A Selva foi criada com a intenção de ser parcialmente revista. Esta parcialidade refere-se não à quantidade de revisão feita, mas sim à qualidade. A idéia é que algumas características sejam lingüisticamente revistas, e que, portanto, a revisão não seja feita árvore a árvore (ou frase a frase), mas caso a caso (diferentemente do *Bosque*, onde todas as frases foram revistas por lingüistas).

Estruturas envolvendo sintagmas nominais, pela freqüência na língua, e pela quantidade de funções em que estão envolvidas, foram as escolhidas para iniciar a revisão, cujo andamento está descrito na página de documentação do projeto. A Selva se subdivide em 3 partes:

A *Selva falada* contém cerca de 100 mil palavras e é composta por transcrições de fala: transcrição de entrevistas e de sessões (debates) parlamentares. As entrevistas são do Museu da Pessoa do Brasil e de Portugal, e os debates parlamentares são da Assembléia da República (PT) e da Assembléia Legislativa da Bahia (BR).

A *Selva literária* é composta por cerca de 100 mil palavras e contém textos literários brasileiros e portugueses do final do século XIX e do início do século XX (cerca de 10.000 palavras por autor), recolhidos na Wikisource, e ainda cerca de 10.000 palavras de literatura contemporânea.

A *Selva científica*, com cerca de 100 mil palavras é, mais propriamente, um *subcorpus* acadêmico-técnico-científico, com textos retirados de teses acadêmicas (do Brasil e de Portugal), de artigos da Wikipédia sobre assuntos relacionados às ciências, como astronomia, biologia, física, geografia, geologia, história, lingüística, zoologia etc., e de documentos do Banco Central Europeu e do Banco Central do Brasil.

Bosque: Parte lingüisticamente revista da Floresta. Contém 190 mil palavras, 9.368 frases, retiradas dos primeiros 1000 extratos (aproximadamente) dos *corpora* CETENFolha (textos do jornal brasileiro *Folha de S. Paulo*) e CETEMPúblico (textos do jornal português Público). Desde 2007, o Bosque vem passando por um segundo processo de revisão, em que foram corrigidas algumas pequenas inconsistências e acrescentadas novas etiquetas. A versão atual é o Bosque 8.0. O Bosque contém alguns novos tipos de etiquetas, que perfazem um total de cerca de 15 mil (instâncias de) novas etiquetas revistas, algumas já mencionadas. As outras etiquetas novas são os "procuráveis": uma classe de etiquetas criada para, por um lado, facilitar uma busca por estruturas complexas, e,

[8] A descrição detalhada de todo o material utilizado na Floresta Sintáctica está na página do projeto.

por outro lado, incluir informação adicional que, embora usualmente presente nos compêndios gramaticais, é de natureza mais semântica (ou discursiva) que sintática. No primeiro grupo, estão etiquetas atreladas aos verbos de orações passivas (com ou sem agente), orações passivas pronominais (passivas com "-se"), verbos de orações substantivas, verbos de orações relativas, verbos de orações sem sujeito explícito e verbos de orações sem sujeito formal. No segundo grupo, estão a tipologia de algumas orações adverbiais (concessivas, causais, conformativas, condicionais, consecutivas, temporais e finais) e os diferentes tipos de partitivos, já comentados.

Em termos quantitativos, o Bosque contém cerca de 12.000 novas etiquetas distribuídas entre os diferentes tipos de procuráveis, 2.000 complementos de nome e 600 n-adj[9]. Uma descrição detalhada de todas as etiquetas do *Bosque*, inclusive em termos quantitativos, está em Freitas e Afonso (2008).

Milhafre: a nova interface de busca em árvores sintáticas

Como foi demonstrado nas seções anteriores, a Floresta contém um material riquíssimo em termos de informação lingüística. Porém, quase tão importante quanto a informação que lá está codificada, é saber como encontrar esta informação – e nem sempre a utilização/manipulação de *corpora* é das tarefas mais simples.

Como já mencionado, a informação devolvida pelo PALAVRAS, e sobre a qual se faz a revisão lingüística (no caso do *Bosque* e da *Selva*), é o formato AD (Figura 1). Não esperamos, contudo, que o usuário da Floresta seja um intrépido desbravador das árvores deitadas (embora seria ótimo se assim o fosse). Pelo contrário, assim como, durante o processo de escolha das opções lingüísticas subjacentes à Floresta, imaginamos usuários com perfis distintos, aqui se passa quase o mesmo. Em primeiro lugar, assumimos que o usuário padrão da interface de busca é o usuário lingüista[10]. Supomos, ainda, que o usuário tem pouca familiaridade com sintaxes/formalizações utilizadas em informática, como expressões regulares. O desafio, aqui, é compatibilizar a complexidade inerente às línguas naturais com um recurso capaz de lidar com essa imensa complexidade de uma maneira razoavelmente simples.

A solução encontrada foi, além da ênfase em uma terminologia que fosse de conhecimento relativamente geral, permitir que as operações que se deseja fazer sobre o *corpus* fossem mostradas em linguagem natural. Isto é, uma vez que toda a informação lingüística da Floresta é codificada em pares função e forma, esses são os elementos (ou tijolos) que deverão poder ser manipulados. As operações são as diferentes maneiras de organizar os elementos, isto é, um sintagma (elemento) que contém (operação) outro sintagma (elemento), uma oração que está no mesmo nível de outra oração, e que por sua vez contém outra oração... Enfim, conseguindo manipular esses dois elementos (forma/função e relações), é possível construir complexas expressões de busca. A Figura 2 mostra a tela inicial do Milhafre, a interface criada para realizar buscas sobre os corpora da Floresta[11].

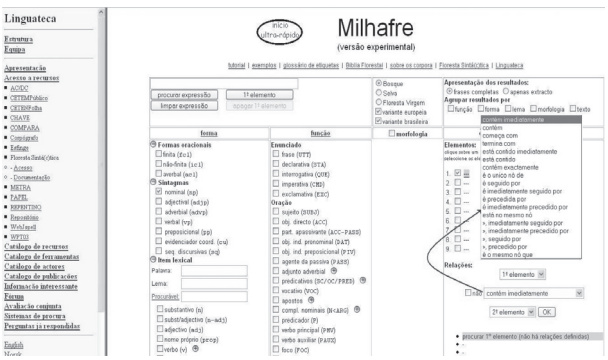

Figura 2. Tela inicial do Milhafre.

A idéia subjacente está em selecionar os elementos (uma forma, uma função, ou ambas) e concatenar esses elementos por meio da seleção das relações disponíveis. Também é possível selecionar um par forma e função, em que o par conta como um único elemento, (por exemplo: sujeito (função) + oração (forma), que pode ser lido como "oração subjetiva"), sem a necessidade de introduzir relações; ou, ainda, apenas uma forma (ou apenas uma função).

Além disso, a interface oferece uma série de outras possibilidades: os resultados podem ser apresentados no formato de concordância, podem conter a indicação de freqüência, podem ser agrupados por lema, por forma ou por função, podem ser visualizados apenas por extrato (em oposição à frase completa), pode-se pedir para que determinados elementos da busca venham sublinhados, pode-se escolher sobre qual material se deseja fazer a busca, bem como sobre qual variante (português do Brasil ou de Portugal), e mais algumas outras funcionalidades. Os resultados podem vir em formato html ou ainda em formato txt, de maneira a facilitar a edição em progra-

[9] Embora as etiquetas descritas no Bosque existam também na Selva e na Amazônia, relembramos que não foram revistas nestes últimos.

[10] Como, em geral, o interesse de engenheiros e informatas está na obtenção de vasto material para treino (e não de apenas determinadas estruturas lingüísticas), supomos (erradamente, talvez) que este não é o nosso usuário padrão.

[11] A Floresta também pode ser interrogada por meio das ferramentas/interfaces Águia e Tgrepeye, embora estes só efetuem buscas sobre versões anteriores do Bosque e da Floresta Virgem.

mas editores de texto, se for necessário. Além disso, a interface dispõe de uma série de exemplos de buscas e de um tutorial.

Considerações finais

A Floresta Sintá(c)tica é um projeto de criação e disponibilização de um *corpus* sintaticamente anotado do português, e neste artigo apresentamos duas novas partes do projeto: *Selva* e *Amazônia*. Manipular (e pesquisar) um grande *corpus* não é simples, e, para facilitar a tarefa, foi construída a interface Milhafre. Como cada uma das partes da Floresta tem diferentes características, é mais adequada para um determinado tipo de aplicação: se o interesse está principalmente em dados quantitativos, a *Floresta Virgem* e a *Amazônia* são o material mais apropriado; se o foco não é tanto a quantidade, mas sim a precisão dos resultados, indicamos o *Bosque* (e, eventualmente, a *Selva*); se a intenção é contrastar determinadas estruturas em diferentes gêneros textuais, a *Selva* é o mais adequado.

De um ponto de vista lingüístico, um dos desafios do projeto está em compatibilizar, de um lado, o usuário lingüista, que pode ter interesse em diferentes modelos teóricos, possuir os mais diferentes graus de conhecimento lingüístico e pouca familiaridade com determinadas formalizações mais utilizadas em informática e, de outro, um único modelo de anotação sintática, subjacente ao *parser* PALAVRAS (Constraint Grammar), modelo freqüentemente pouco conhecido do lado "linguístico não-computacional" e uma interface de acesso a manipulação de *corpus* capaz de lidar com um objeto altamente complexo como a língua.

A solução proposta consiste em, ao lado de uma terminologia razoavelmente aceita, buscar uma análise lingüística que reflita uma perspectiva mais consensual e que permita a maior possibilidade de abordagens sobre um dado fenômeno. Ou seja, procuramos não entrar na questão de se uma estrutura como "3 das crianças" é um partitivo ou "apenas" um quantificador, mas tentamos deixar o campo aberto para pesquisas.

Referências

AFONSO, S.; BICK, E.; HABER, R.; SANTOS, D. 2001. Floresta sintá(c)tica: um treebank para o português. *In*: ENCONTRO DA ASSOCIAÇÃO PORTUGUESA DE LINGUÍSTICA, XVII, Lisboa, 2001. *Actas*... Lisboa. Disponível em http://www.linguateca.pt/Diana/download/AfonsoetalAPL2001.rtf .Acesso em 16/12/2008.

AFONSO, S. 2004. A Floresta Sintá(c)tica como recurso. Disponível em http://www.linguateca.pt/documentos/Afonso2004Recurso.pdf. Acesso em 16/12/2008

AZEREDO, J.C. 2001. *Iniciação à Sintaxe do Português.* Rio de Janeiro, Jorge Zahar Ed., 172 p.

BASÍLIO, M. 2004. *Formação e classe de palavras no português do Brasil*. São Paulo, Contexto, 96 p.

BICK, E. 2000. The Parsing System PALAVRAS: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. Dinamarca. Tese de doutorado. Aarhus University.

FRANCHI, C. 2006. *Mas o que é mesmo "gramática"?* São Paulo, Parábola, 151 p.

FREITAS, C.; AFONSO, S. 2008. Bíblia Florestal: Um manual lingüístico da Floresta Sintá(c)tica. Disponível em: http://www.linguateca.pt/Floresta/BibliaFlorestal/. Acesso em: 16/12/2008.

JI, H.; GRISHMAN, R. 2006. Data Selection in Semi-supervised Learning for Name Tagging, *In*: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, Sydney, 2006. *Anais*... Sydney, p. 48-55.

MEYER, R.M.B. 1995. A questão do complemento nominal. *In*: J. HEYE (org.), *Flores Verbais: uma homenagem lingüística e literária para Eneida do Rego Monteiro Bomfim no seu 70o aniversário*. Rio de Janeiro, Ed. 34, p. 161-176.

PERES, J.A. 1992. Questões de Semântica Nominal. *Cadernos de Semântica 1*, p. 1-35.

PERINI, M.A. 1986. *Para uma nova gramática do português*. São Paulo, Editora Ática, 94 p.

PERINI, M.A. 1997. *Sofrendo a gramática.* São Paulo, Ática, 104 p.

Submetido em: 16/10/2008
Aceito em: 05/11/2008

Claudia Freitas
Universidade de Coimbra, Pólo de Coimbra da Linguateca - DEI
Coimbra, Portugal

Paulo Rocha
Universidade de Coimbra, Pólo de Coimbra da Linguateca - DEI
Coimbra, Portugal

Eckhard Bick
Universidade do Sul da Dinamarca
Odense, Dinamarca