Revisão humana da Floresta Sintá(c)tica: exemplos e método

Raquel Marchi – Linguateca
Novembro de 2004

1. Introdução

Devido à complexidade da linguagem humana, como em casos de ambigüidade, anáfora, elipse, etc... somente uma análise automática das sentenças pode apresentar, em seu resultado, um grande número de análises incorretas (ruído). Por este motivo, para uma análise mais confiável, se faz necessária a revisão – seguida de correção, sempre que for o caso – por um especialista humano (revisão manual) das sentenças sintaticamente analisadas.

Assim como em o *CETEMPúblico* (vertente portuguesa da Floresta Sintá(c)tica – conjunto de sentenças sintaticamente analisadas (manualmente revisadas ou não) para a língua portuguesa, comumente chamado na Lingüistica Computacional de "*treebank*"), após a análise automática, e sempre com o intuito de reduzir ruídos, também parte das sentenças do *CETENFolha* – parte brasileira da Floresta Sintá(c)tica, composta por extratos de textos jornalísticos do jornal brasileiro Folha de São Paulo – foram e ainda estão sendo submetidas a revisão humana.

Os extratos que compõem a Floresta Sintá(c)tica foram escolhidos aleatoriamente dos cadernos que compõem os jornais (Público e Folha de São Paulo) e processados/analisados automaticamente por um analisador sintático – o *parser PALAVRAS*[1] e a revisão humana (tanto para a parte européia, como para a brasileira) é feita por especialistas lingüistas e foi dividida em 2 etapas:

- revisão das sentenças analisadas no formato *C.G.* (*Constraint Grammar*);
- e revisão no formato de árvores de constituintes (comumente chamada pelo grupo de "árvores deitadas").

A seguir, nas próximas seções, falaremos brevemente destas 2 etapas de revisão humana para o *CETEN Folha*[2] – seguida da apresentação de alguns exemplos de casos de ruído encontrados nessa revisão – e da importância desse tipo de *corpora* manualmente revisado no Processamento de Linguagem Natural (PLN) – e consequentemente do papel da Floresta Sintá(c)tica para o avanço da Lingüística Computacional para o português.

2. Revisão humana do CETENFolha

A parte manualmente revisada recebeu o nome de ***Bosque*** e a revisão humana do *CETENFolha* conta com a participação de um revisor nativo do português do Brasil e que também já era familiarizado com a revisão de sentenças da parte do *CETEMPúblico*.

---

[1] Bick, Eckhard (2000). *The Parsing System Palavras, Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Århus University Press.

[2] **http://www.linguateca.pt/CETENFolha/**

Em ambas etapas, as sentenças recebem a etiquetação (*tagging*) de acordo com a simbologia utilizada pelo *parser PALAVRAS*[3] - e obedecem todos os critérios relativos à documentação[4] já existente sobre a revisão humana do *CETEMPúblico*. E são feitas sentença a sentença. As dúvidas ou algumas considerações surgidas durante a revisão são reportadas periodicamente ao grupo e discutidas em conjunto, para depois, serem devidamente documentadas.

2.1 Ruído

Como já foi dito, a presença de ruído é inevitável no resultado da análise automática. Construções da língua podem gerar ambigüidade ou estruturas complexas, o que dificulta a análise (em alguns casos dificulta até a análise humana).

De um modo geral, é possível separar as incorreções encontradas, decorrentes da análise automática, basicamente em dois grupos:

- aquele em que as incorreções são fruto de algum tipo de ambigüidade na classificação morfológica e/ou sintática das palavras, e que são, de alguma forma simples de resolver com a intervenção humana;
- e outro em que as incorreções são fruto de estruturas complexas da língua, de resolução difícil até para o revisor humano (tais casos requerem discussão e opinião de todo o grupo envolvido no projeto).

2.1.1 Ambigüidade morfológica e sintática

O primeiro grupo de incorreções se faz, talvez em parte, pela falta de informação semântica nas palavras que compõem as sentenças. Muitas palavras podem receber classificação diferente, dependendo do contexto em que foram inseridas em uma determinada sentença. Os exemplos mais comuns deste tipo de ambigüidade, são os de palavras que podem ser classificadas tanto como substantivo, quanto adjetivo (palavras como "*brasileiro/a*"; "*jovem*"), ou ainda, como adjetivos e advérbios ("*rápido/a*", por exemplo) e adjetivos e particípio passado de alguns verbos ("*passado/a*"; "*eleito/a*", etc.). É importante dizer que, às vezes, essa ambigüidade pode apresentar mais de uma leitura de um determinado sintagma ou sentença, e nestes casos a solução encontrada durante a revisão humana é a de trazer uma análise de acordo com cada leitura. Um exemplo seria um sintagma nominal do tipo "*O jovem trabalhador*", que permite a leitura de "*jovem*" ora como modificador do núcleo do sintagma, ora como sendo o próprio núcleo, estando ambas corretas.

O contexto também é importante no papel/função que cada palavra desempenha na sentença como um todo, quando é preciso decidir se, em uma sentença, aquela palavra ou grupo de palavras é o sujeito ou o objeto/predicativo do sujeito na oração, por exemplo. Muitas vezes, a ordem dos elementos na sentença pode confundir a análise automática.

```
STA:fcl
=ACC:fcl
==SUBJ:np
===>N:adj('novo' F S)      Nova
```

[3] **http://visl.sdu.dk/visl/pt/info/symbolset-floresta.html**

[4] Afonso, Susana e Ana Raquel Marchi. **Critérios de separação de sentenças/frases**
Afonso, Susana e Ana Raquel Marchi. **A etiqueta <sic> </sic>**

```
===H:n('droga' F S) droga
==P:v-fin('combater' PR 3S IND)  combate
==ACC:n('asma' F S) asma
=,
=P:v-fin('dizer' PR 3S IND)      diz
=SUBJ:n('estudo' M S)     estudo              #ACC
```

No exemplo acima, com a reordenação da sentença, tem-se: "*Estudo diz que nova droga combate asma*", em que a palavra "*estudo*" é o núcleo do sintagma nominal (sujeito da oração) e não seu objeto direto, como havia sido anteriormente analisado (*#ACC*).

A locução preposicional "*além de*", classificada como parte de uma estrutura adverbial, também pode, em alguns casos, desempenhar o papel de conjunção coordenativa, assumindo valor de "e" na oração. O mesmo vale para "*como também*". Seguem alguns exemplos.

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

CF64-1 Para tanto é imprescindível priorizar algumas ações como a recuperação
de unidades armazenadoras de cereais e das rodovias, além do incentivo à
utilização das ferrovias.

CF97-2 No estande da Tec Toy há três consoles Mega Drive, além de um Master
System e quinze cartuchos de jogos.

CF186-3 Inicialmente, a equipe econômica chegou a estudar a privatização das
empresas do grupo Nuclebrás, além de Angra 1 e 2, mas não chegou ao fim.

CF346-3 No PPR, a discussão será sobre a política de alianças e se ela inclui
o PSDB, além dos novos candidatos do partido a presidente.

CF225-1 Elas não apenas restringem a liberdade e o futuro dos que tentam
deixar Cuba como também oferecem expectativa de «reformas democráticas
forçadas» aos que ficam na ilha.
```

De forma resumida, as palavras podem ser ambíguas na morfologia, ou seja, na sua classificação individual (uma mesma palavra pode pertencer a classes gramaticais diferentes), ou na sintaxe, no papel ou função que ela desempenha dentro da sentença. Pois dependendo da sentença, até um simples "determinante" em um sintagma nominal, como o artigo, pode ocupar papel de núcleo deste. O exemplo seriam sentenças do tipo, "*O garoto da direita havia feito a pergunta e o da esquerda devia respondê-la*". Na primeira oração, o sujeito "*o garoto da direita*" tem como núcleo a palavra "*garoto*". Na segunda oração, em vez, o núcleo do sujeito "*o da esquerda*" é "*o*", visto que a palavra "*garoto*" está implícita (elíptica) nesta oração. Neste caso, a palavra "*o*", artigo definido no masculino singular, vai ter sempre a mesma classificação morfológica nas duas orações, mas assumirá diferentes papéis sintáticos (ora determinante, ora núcleo).

Nestes casos, a intervenção humana, ciente do contexto de cada palavra inserida na sentença, pode resolver facilmente aquilo que pode não ter sido resolvido no processamento automático.

2.1.2 Estruturas complexas

As estruturas complexas são aquelas de difícil representação no formato de árvores de constituintes.

O caso de elipse, como foi visto de forma bastante simples no exemplo de "*o garoto da direita*" e "*o da esquerda*", pode representar um problema de difícil resolução e representação quando em estruturas mais complexas, ou seja, quando por exemplo um dos constituintes, em uma estrutura coordenada, é elíptico em uma das orações.

```
CF144-5 Para outros 9%, haverá resistências mas não (será) a ponto de
inviabilizar a posse.

CF201-1 Aristides está esperando a publicação do texto do FSE aprovado pelo
Congresso para concluir se a desvinculação dessas verbas atinge ou não
(atinge) direitos individuais.

CF215-2 Parágrafo 4º -- As contribuições para a Seguridade Social, de que
tratam os arts. 20, 21, 22 e 24 da Lei nº 8.212, de 1991, serão convertidas
em URV e convertidas em UFIR nos termos do art. 53 da Lei nº 8.383, de 30 de
dezembro de 1991, ou (serão convertidas) em cruzeiros reais na data do
recolhimento, caso este ocorra antes do primeiro dia útil do mês subsequente
ao de competência.

CF314-3 «(Sou)Insana, mas (sou) saudável»

CF386-3 O lucro cresceu 24% para US$ 51 milhões e as vendas, 18%, (cresceram)
para US$ 776 milhões.
```

Como na maioria dos casos, é o verbo o elemento implícito na oração, fica difícil representar, nas estruturas das árvores, que todos os outros elementos de um predicado verbal continuem ali, com as mesmas funções.

Além dos casos complexos de coordenação com elementos implícitos, durante a revisão humana das sentenças do *CETENFolha*, outros tipos de coordenações se mostraram bastante complexas em diferentes níveis de construções. Elas podem compartilhar, por exemplo, o mesmo sujeito, mesmo adjunto adverbial ou predicativo e até mesmo verbo auxiliar. Seguem alguns exemplos:

```
CF8-9 Se eu dirigisse uma federação, (eu) apresentaria balanços mensais e
(eu) liberaria minhas contas bancárias.

CF147-1 Outro lado de Lara: ela estudou canto lírico, adora música e costuma
cantar em shows e jam sessions com amigos.

CF298-5 Ai, mãe Menininha, acode-nos nesta hora de quase desespero, dá-nos o
alimento da confiança e do sonho.
```

Cada oração coordenada, nos exemplos acima, compartilham o mesmo sujeito. Um exemplo é o da sentença CF147-1, "*(Lara) ela estudou canto lírico; (Lara) adora música e (Lara) costuma cantar em shows e jam sessions com amigos*". Neste e nos outros exemplos também, cada predicado apresenta verbos com seus próprios complementos e objetos. Como vimos, "*Lara adora música*" e "*Lara costuma cantar em jam sessions*".

```
CF101-2 Lula cresceu e prosperou enquanto tudo dava errado no país.
```

CF182-2 Se eu fosse você, passava a andar de táxi especial e apresentava a conta à concessionária.

CF210-3 Preocupado com a facilidade de comunicação de seu adversário, Francisco Rossi, Mário Covas acatou as recomendações de assessores do presidente eleito, Fernando Henrique Cardoso, e se submeteu às técnicas de marketing.

No exemplo CF101-2, acontece o mesmo fenômeno citado acima, com a diferença de que o fato de "*Lula ter crescido e prosperado*" é contemporâneo ao fato de que "*tudo dava errado no país*", no caso uma oração coordenativa adverbial temporal. Ou simplesmente as sentenças podem apresentar também outros complementos em comum, sejam advérbios ou objetos, como pode ser visto nos outros exemplos.

CF193-5 À tarde, vai a Belo Horizonte e visita o jornal «O Estado de Minas» em companhia de Hélio Costa, candidato do PP ao governo do Estado.

Na sentença CF193-5, as ações "ir" ("*vai a Belo Horizonte*") e "visitar" ("*visita o jornal...*") compartilham o mesmo adjunto adverbial, ou seja, acontecem no mesmo período de tempo, "*à tarde*". Diferente da sentença CF114-5 – "*A ação da PF começou às 16h30 de anteontem e só terminou ontem às 15h30.*" – em que a ação começa num determinado período de tempo e termina em outro.

Para os casos de sujeito partilhado, considera-se que o elemento compartilhado permanece fora da coordenação e opta-se por uma estrutura subespecificada (representada pelo símbolo de interrogação "?"), a qual recebe uma etiqueta secundária de predicado (<predicate>), representando a estrutura que engloba o verbo e seus complementos e/ou adjuntos.

```
=====SUBJ:pron-pers('ela' F 3S NOM)     ela
=====?:cu
======CJT:? (<predicate>)
=======P:v-fin('estudar' PS 3S IND)     estudou
=======ACC:np
========H:n('canto' M S)   canto
========N<:adj('lírico' M S)     lírico
======,
======CJT:? (<predicate>)
=======P:v-fin('adorar' PR 3S IND)      adora
=======ACC:n('música' F S) música
======CO:conj-c('e')       e
======CJT:? (<predicate>)
=======P:vp
========AUX:v-fin('costumar' PR 3S IND) costuma
========MV:v-inf('cantar') cantar
=======ADVL:pp
========H:prp('em') em
========P<:cu
=========CJT:n('show' M P) shows
=========CO:conj-c('e' <co-prparg>)     e
=========CJT:n('jam_sessions' F P)      jam_sessions
=======ADVL:pp
========H:prp('com')       com
========P<:n('amigo' M P)  amigos
```

ou...

```
STA:fcl
=SUBJ:prop('Lula' M S)     Lula
=?:cu
==CJT:? (<predicate>)
===P:v-fin('crescer' PS 3S IND)  cresceu
==CO:conj-c('e')    e
==CJT:? (<predicate>)
===P:v-fin('prosperar' PS 3S IND)prosperou
=ADVL:fcl
==ADVL:adv('enquanto' <rel> <ks>)enquanto
==SUBJ:pron-indp('tudo' <quant> M S)    tudo
==P:v-fin('dar' IMPF 3S IND)     dava
==ACC:v-pcp('errar' M S)   errado
==ADVL:pp
===H:prp('em' <sam->)      em
===P<:np
====>N:art('o' <-sam> <artd> M S)o
====H:n('país' M S) país
=.
```

Um outro exemplo de estrutura complexa, mas que não envolve coordenação, é o que chamaremos de “parênteses” (mesmo se em alguns casos, ele fisicamente não exista), palavra que nesse caso indica não só a presença deste sinal de pontuação, mas que também introduz um comentário (observação, explicação, de diálogo com o leitor) que, geralmente é externo à sentença. A seguir, veremos alguns exemplos recolhidos do *corpus* do *CETENFolha*.

```
CF150-5 Exercendo o direito de não estar a favor ou contra ninguém, gostaria
aqui de remar contra a maré e pasmem manifestar meu otimismo.

CF251-2 No caso do autor de «O Amanuense Belmiro» (recomendo a leitura de um
dos inaugurais romances urbanos brasileiros) não havia nada especial a
demandar a pergunta feita através do Alcino.
```

Nestes casos, considera-se a sentença entre parênteses como um enunciado que está dentro de outro.

```
STA:fcl
=ADVL:pp
==H:prp('em' <sam->)       Em
==P<:np
===>N:art('o' <-sam> <artd> M S)  o
===H:n('caso' M S)  caso
===N<:pp
====H:prp('de' <sam->)     de
====P<:np
=====>N:art('o' <artd> <-sam> M S)      o
=====H:n('autor' M S)      autor
=====N<:pp
======H:prp('de')   de
======"«
======P<:prop('O_Amanuense_Belmiro' M S)       O_Amanuense_Belmiro
======"»
=(
=STA:fcl
==P:v-fin('recomendar' PR 1S IND)recomendo
==ACC:np
===>N:art('o' <artd> F S)  a
===H:n('leitura' F S)      leitura
===N<:pp
```

```
====H:prp('de')     de
====P<:np
=====H:num('um' <card> M S)      um
=====N<:pp
======H:prp('de' <sam->)   de
======P<:np
=======>N:art('o' <-sam> <artd> M P)    os
=======>N:adj('inaugural' M P)   inaugurais
=======H:n('romance' M P)  romances
=======N<:adj('urbano' M P)      urbanos
=======N<:adj('brasileiro' M P)  brasileiros
=)
=ADVL:adv('não')    não
=P:v-fin('haver' IMPF 3S IND)    havia
(...)
```

Estes exemplos ilustram um pouco dos problemas encontrados no processo de revisão manual/humana do resultado da análise automática das sentenças que compõem a Floresta Sintá(c)tica devido a complexidade das línguas naturais.

3. Importância de um *corpus* manualmente revisado para o português

Grandes *corpora* analisados como do projeto o Penn Treebank, Susanne, BNC (*British National Corpus*)[5] são importantes para a lingüística de *corpus*, principalmente na construção de recursos lingüísticos como léxicos e gramáticas, e também para a Lingüística Computacional, na aquisição e avaliação de ferramentas de PLN, como os analisadores automáticos (*parsers*), ou ferramentas de tradução automática (visto que apresentam um vasto exemplário de padrões estruturais da língua).

Apesar da importância, até hoje, a maioria dos projetos existentes são para o inglês. Por esse motivo, a construção de um *treebank* para o português (europeu e do Brasil) - a Floresta Sintá(c)tica – representa um grande avanço para o PLN nessa língua, por oferecer a matéria-prima para todas essas aplicações.

A Floresta Sintá(c)tica representa fonte de dados e padrões estruturados de linguagem para o português europeu e do Brasil que, além de base para o desenvolvimento e treinamento de *parsers* e outras ferramentas para o processamento de linguagem, podem ser utilizados também para outras aplicações:

a) como já foi dito, na **construção de léxicos e gramáticas** (pertencentes a um domínio específico ou não) da língua portuguesa;
b) **aquisição e/ou extração de terminologia ou de polilexicais** (ou "*multiwords*" – que são grupos fixos ou semi-fixos de palavras que ocorrem com uma determinada freqüência em uma determinada língua e que, muitas vezes, têm significado somente se juntas e que são atualmente objeto de inúmeros estudos, projetos[6] e *workshops*[7] para variadas línguas) através de regras gramaticais utilizando padrões sintáticos;

[5] **http://treebank.linguist.jussieu.fr/**
[6] CSLI Linguistic Grammars Online (LinGO) Lab at Stanford University.
http://mwe.stanford.edu/
[7] ACL 2003 *Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*
http://www.cl.cam.ac.uk/users/alk23/mwe/mwe.html

c) pode ser utilizada no **estudo de fenômenos lingüísticos**, como uso de preposições ou mesmo estruturas de coordenação e subordinação, entre outros, que poderiam, entre outras coisas, contribuir para as aplicações já citadas;
d) ou **estudos comparativos** desses fenômenos ou estruturas sintagmáticas ou sentenciais entre vertentes de uma mesma língua, como é o caso de uma comparação entre o português europeu e o brasileiro, ou ainda do português com uma ou mais línguas.

Tudo isso vem ilustrar, como já foi dito, a importância do projeto Floresta Sintá(c)tica para o tratamento computacional do português europeu e brasileiro.

Agradecimento
A Floresta Sintáctica é (parcialmente) financiada pela Fundação para a Ciência e Tecnologia, co-financiada pelo POSI, através do projecto POSI/PLP/43931/2001 (Linguateca).