Floresta Sintá(c)tica: apresentação e história do projecto

Diana Santos, Eckhard Bick e Susana Afonso
www.linguateca.pt/Floresta/

História

- **2000** Início como continuação de uma boa colaboração no AC/DC entre o VISL e o ProcCompPort
- **2000-2001** Um ano de trabalho activo para plantar as raízes: três bolseiros de linguística em Odense, formação, criação do método de trabalho, escolha e limpeza do texto, primeiras ferramentas computacionais em Oslo
- **2002-2004** Mais alguns anos com trabalho parcial das anteriores bolseiras, aumentando o número de árvores e a quantidade de fenómenos revistos
- Treino de um novo membro da equipa de linguística
- **2005** Validação da Floresta pelo pólo de Braga, vários novos formatos
- Floresta adormecida...

História (presente)

- Em 2007: nova equipa, nova localização
  - Liderança: Eckhard Bick
  - Cláudia de Freitas: responsável pela parte linguística, em Coimbra
  - Paulo Rocha: responsável pela parte informática, em Coimbra
- Período de transição e formação na Floresta dos novos membros
  - Junho de 2007 ao presente
  - encontro inicial em Oslo, Junho de 2007
- Pontapé de saída e passagem de testemunho oficial: o presente encontro

O que é a Floresta Sintá(c)tica?

- Um recurso
- Um projecto
- Um serviço
- Um estado de espírito

A FS como recurso

- Um conjunto de conjuntos de frases (“extractos” do CETEMPúblico e do CETENFolha)
- anotados automaticamente com informação gramatical: morfológica, sintáctica
- revistos por linguistas segundo um conjunto de directivas que vão sendo criadas e tornadas públicas (a chamada bíblia florestal)

- Existem vários formatos da Floresta, para simplificar o seu uso por uma população variada, neste momento todos equivalentes
  - CG(D), AD, VISL AD, TIGER, ADCS, ...

Alguns dados sobre a FS como recurso

Orações	21.931
Orações finitas	15.566
Orações infinitivas	5.602
Orações averbais	763
Sintagmas* nominais	43.096
Sintagmas* preposicionais	32.210
Sintagmas* adjectivais	1.780
Sintagmas* adverbiais	833
Itens coordenados	5.448
Árvores	9.431

** mais do que uma palavra*

A FS como projecto

- Primeiro, árvores criadas pelo PALAVRAS
- Depois, revisão dessas árvores
- Mais árvores, mudanças ao PALAVRAS
- Depois, revisão dessas árvores e das anteriores

(A revisão inclui anotação de coisas que o PALAVRAS não sabe)

- Validação da FS a nível sintáctico (sintaxe da FS)

- Versões criadas regularmente
- Disponibilizadas directamente na rede
- Colocadas acessíveis através de serviços na rede: Águia, CorpusEye...

A FS como serviço

- Um sistema de procura (ou vários) que permite(m) interrogar o recurso da Floresta
- Um recurso posto ao serviço da comunidade

- Um sítio onde se podem fazer procuras complexas e obter texto e/ou distribuição
- Um sítio onde se podem fazer procuras simples e obter árvores (objectos) complexos

- Um lugar onde se pode ir buscar texto muito anotado para trabalhar

A FS como estado de espírito

- Curiosidade sobre a maneira como o português funciona
- Humildade perante a criatividade dos falantes
- Desejo de servir a comunidade que trabalha no proc. comp. do português

- Interesse em dar novos mundos ao mundo (da linguística ou do processamento de linguagem natural)
- Fornecer instrumentos e hipóteses
- Dialogar com os outros interessados

A Floresta no estrangeiro

- Usada por Sabine Buchholz & Darren Green num artigo no LREC 2006 para ilustrar problemas de manutenção de uma floresta
- Usada por Jason Balridge para inferir uma gramática do português
- Usada pela "tarefa partilhada" do CoNLL-X 2006, *ConLL-X shared task on multilingual dependency parsing*
- Integrada por Steven Bird em Setembro de 2007 no NLTK, *Natural Language Toolkit*
- Outros pedidos
  - John Hopkins University
  - Essex University
- Distribuição da origem dos pedidos (sem comunicação)

Um exemplo

CF185-7 Ele sequestrou e violentou três meninos com a intenção de lhes transmitir o vírus da Aids de que se sabia portador.

```
======>MV:v-inf('transmitir')                transmitir
======>ACC:np
=======>N:art('o' <artd> M S)                o
======>H:n('virus' M S)      virus
======>N<:pp
=======>H:prp('de' <sam->)                   de
========P<:np
========>N:art('o' <-sam> <artd> F S)   a
========H:n('aids' F S)      Aids
========N<:fcl
==========OC:adjp-
==========N<:pp
===========H:prp('de')    de
===========P<:pron-indp('que' <rel> M S)              que
==========ACC:np
==========H:pron-pers('se' M 3S ACC)  se
==========P:vp
==========MV:v-fin('saber' IMPF 3S IND)               sabia
===========OC:adjp
===========H:adj('portador' M S)         portador
```


Quantas questões esta frase ilustra?

- orações relativas
  - quantos tipos?
- coordenação
  - partilha de argumentos: *três meninos* é objecto de *sequestrou*; *Ele* é sujeito de...
- ligação vaga (2 vezes)
  - *portador do vírus da Aids* ou *portador de Aids*
  - *sequestrou e violentou* ou só *violentou (?) com a intenção de...*
- constituintes descontínuos (ou dependências de longa distância)
- elipse? *saber-se portador*
- *Aids* é um nome próprio?

O quebra-cabeças é...

- integrar todas as peças de forma consistente numa mesma árvore
- ao contrário da maior parte dos projectos e/ou estudos de corpora que só observam um fenómeno de cada vez
- cada frase é geralmente um exemplo de tantos fenómenos quantos sintagmas ou palavras (generalização apressada ☺)
- e ainda há o problema do léxico – o que são palavras ou locuções ou morfemas
  - ex-comandante da LUAR

O mito da neutralidade e da facilidade

- Quer-se um recurso fácil de compreender por qualquer pessoa / qualquer linguista
- uma interface "intuitiva"
- para um objecto com uma enorme complexidade
- é um paradoxo!
- não será pedir demais?
- o único paradigma que pode funcionar é o "procurar por exemplos", tentativa e erro

Duas acções distintas na procura em corpora

- procurar exemplos (**concordâncias**)
  - casos de objectos directos com uma oração relativa
  - *troquei a faca que o meu tio me deu*
- fazer consultas agregadas (**distribuição**) (lista e frequência)
  - lista de verbos que têm objectos com uma oração relativa (*trocar*)
  - lista de verbos que ocorrem na oração relativa que faz parte de um objecto directo (*dar*)
  - lista de sujeitos que ocorrem nas orações relativas que fazem parte dos objectos directos (*o meu tio*)
  - lista de assinaturas da oração relativa (*pron-rel np pron v-fin*)
  - lista de assinaturas de função da oração relativa (*ACC SUJ DAT P*)

Alguns tipos de procuras à Floresta

- Procuras simples de objectos
  - orações relativas, sintagmas adjectivais, apostos, pronomes clíticos
- Procuras aos níveis dos constituintes directos
  - orações subordinadas com verbo no conjuntivo
  - sintagmas nominais com núcleo adjectival
  - orações relativas introduzidas por um advérbio
  - frases com três sintagmas preposicionais
- Procuras ao nível dos co-ocorrentes directos
  - complementos nominais de haver
  - verbos com objecto frásico
  - verbos usados reflexivamente

Mas a classificação não é óbvia!

- Em última análise, a forma como a distinção está codificada em qualquer floresta é arbitrária! (**mas pode ser resolvida com um sistema de procura adequado**)
- `icl`, `acl`, `fcl` existem, mas `rcl` ou `subcl` não existem
- O tempo está marcado no verbo, ou só/também na oração?
- O género está marcado no SN, ou só no seu núcleo, ou em cada palavra passível de ter género?
  - *um índio pele vermelha* : que género deve ser marcado em *vermelha*? e em *pele*? e em *pele vermelha*?
- 3/4 razões para ter um adjectivo como núcleo
  - elipse, propriedade, indeterminação: *jovens alemães*

E as necessidades do utilizador não são precisas!

- sintagma nominal *é quando* sintagma nominal
- SN complexos, mas sem oração no meio
- verbos que aparecem após uma citação
- frases em que a ordem sujeito-verbo-objecto é quebrada
  - frases que têm os 3, e verbo sem auxiliares?
  - VSO, SOV, OSV, OVS, VOS
- encontre um SN com a maior quantidade possível de dependentes
  - pai [de família [de emigrantes [dos subúrbios [de Moscovo [de 1900]]]]]
  - cão [de caça] [de loiça] [da Bélgica] [do meu pai] [do tempo da Grande Guerra]
- orações em que o particípio não exerce uma função verbal

Em conclusão

- O que é complexo não pode ser reduzido ao simples para quem não percebeu a complexidade
- O que é complexo exige conhecimento e aprendizagem, não tem UMA resposta simples
- a interacção com a Floresta Sintá(c)tica não pode aprender-se numa hora
- sem perceber as distinções sintácticas feitas pela língua portuguesa, e a(s) forma(s) como elas foram codificadas pela equipa, não se pode interrogar a Floresta
- estamos (e sempre estivemos!) dispostos a dialogar e a explicar melhor as centenas de opções tomadas, e a florestar em conjunto