Florestas (*treebanks*)
apresentação geral, com especial relevo para a Floresta Sintá(c)tica

Susana Afonso
Diana Santos

Plano

- **Treebanks e Floresta Sintá(c)tica**
  - Apresentação e história do projecto
  - Usos de uma floresta sintáctica
- **A Floresta vista de dentro**
  - Fases do projecto
  - Opções e desafios
- **A Floresta vista de fora**
  - Vantagens na investigação linguística
- **(Há) Futuro (?)**

Intermezzo 1

- Apresentação rápida de
  - Floresta Sintá(c)tica: apresentação e história do projecto (parte1)
  - Aplicações de uma floresta sintáctica

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)

Aplicações de uma floresta sintáctica

Diana Santos
www.linguateca.pt

Para que serve

- qual a função
- qual a utilidade
- qual o resultado/impacto
- quais as consequências

- Primeiro, uma abordagem descritiva
- Depois, uma abordagem crítica

Para ilustrar uma (teoria da) gramática

- Uma coisa é ter uma teoria que descreve uma (ou todas as) línguas
- Outra coisa muito diferente é
  - ter uma aplicação dessa teoria que cobre **texto real**
  - por oposição a texto fabricado para exemplificar uma teoria
- É diferente porque
  - uma gramática (ou teoria da gramática) não especifica geralmente como chegar a um dado resultado – operacionalização
  - o texto tem sempre um número crescente de pormenores – não se pode falar de uma gramática completa (no sentido de que todos os fenómenos já foram descritos)
  - não se pode analisar parcialmente uma frase/não se costuma

Para criar dados para futuro processamento

- se se conseguir obter um número significativo de casos
- pode-se desenvolver modelos (ou treinar sistemas) que usam esses casos para analisar mais texto
- pode-se criar regras ou hipóteses para exploração linguística mais detalhada em corpora maiores
  - se a maior parte dos casos de comparação representam ironia, pode-se usar essa regularidade para extrair grandes quantidades de candidatos a casos de ironia de corpora
  - se a maior parte das orações relativas explicativas aparecem associadas ao sujeito, pode-se usar essa "regra" para extrair candidadtos a sujeitos
  - se a coordenação de adjectivos é frequentemente usada após sintagmas preposicionais em alguns casos e não noutros…

Para avaliar sistemas

- desenvolvidos independentemente
  - comparando os resultados
  - obtendo dados específicos
  - comparando abordagens de anotação (estudos de mutilação)
- desenvolvidos ou melhorados durante o processo
  - o PALAVRAS
  - sistemas de REM
  - sub-sistemas do AC/DC (atomização e separação de palavras)
- para fazer uma avaliação conjunta em sintaxe computacional
  - para obter dados aos quais já está associada uma distribuição
  - para comparar tarefas específicas sobre as quais há ou é possível chegar a um consenso

Para fazer investigação em sintaxe... e semântica

- "mãos na massa"
- dado um conjunto de frases sistematicamente analisadas e compreendidas pelos falantes
- comparar com as intuições
- descobrir casos complicados
- estudar a interacção de fenómenos pertencentes a esferas diferentes
- identificar casos excepcionais
- ensinar sintaxe

Para fazer investigação em informática

- Que tipos de formalismos são melhores para descrever o resultado
- Que tipo de gramáticas são necessárias
- Qual o melhor sistema para indexar e validar a informação
- Que tipo de necessidades têm os utilizadores de uma floresta
- Que tipo e forma de resultados são preferíveis
- Sistemas de ajuda à revisão da anotação
- Sistemas de visualização

Descrição de casos conhecidos

- Penn Treebank
  - indução de gramáticas
  - avaliação de análise sintáctica: ParsEval
- SUSANNE
  - descrição minuciosa da língua
  - avaliação de análise sintáctica: GR-scheme
  - criação e avaliação de novas medidas: LAM
- Czech TD
  - construção e melhoria de dicionários
- NEGRA
  - novas metodologias de anotação e criação

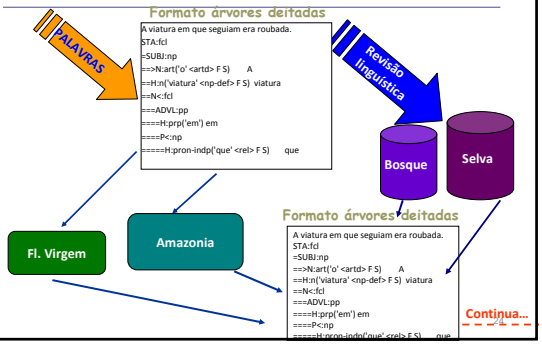
Crítica

- A maior parte das pessoas que usam as florestas não têm a noção do trabalho que lá está incluído, nem do que ainda falta ser feito
- A maior parte das pessoas que trabalham com florestas passam o tempo a criá-las ou a melhorá-las, não a usá-las
- As florestas são um investimento para o futuro, mas geralmente não acompanhado:
  - ainda não existem os utilizadores
  - os futuros utilizadores muito raramente exprimem os seus desejos e/ou necessidades (e quando os exprimem, precisam sempre de corpora muitíssimo maiores!)
  - os gramáticos (fora da equipa) estão aparentemente completamente desinteressados na existência de uma floresta sintáctica ou não para a sua língua

Floresta Sintá(c)tica: mesmo recurso, língua diferente?

- Sim e não
- O que podemos encontrar na Floresta?
  - Variantes: português europeu e do Brasil
  - *Bosque*: totalmente revisto distribuído em vários formatos; género jornalístico (X palavras, Y árvores)
  - *Selva*: parcialmente revisto (caso a caso); vários géneros: selva falada, científica e literária (X palavras, Y árvores)
  - *Amazónia*: não revisto (x palavras, Y árvores)

A Floresta vista de dentro (Freitas 2009)



Floresta Sintá(c)tica vista de dentro

- Duas fases do projecto:
  - **1ª Fase (2000-2005)**, início do Bosque (revisão árvore a árvore) – fase de "desbravamento"
    - Desafios e opções :
    a) Pontuação: *a que nível?*
    b) Expressões multipalavra: *uma só palavra vs. estrutura interna?*
    c) coordenação com argumentos partilhados: *representação sintáctica da partilha?*
    d) estruturas com séries de preposições (*de X a Y*): *semelhante à coordenação?*
    e) Granularidade das etiquetas : *representação sintáctica de diferenças semânticas?*
    f) Vagueza estrutural: *uma vs. mais alternativas?*
    g) ...

Pontuação: a que nível?

CP1-3 É uma das mais antigas discotecas do Algarve, **situada em Albufeira,** que continua a manter os traços decorativos e as clientelas de sempre
A1
STA:fcl
=P:vp
==MV:v-fin('ser' <nosubj> PR 3S IND) É
=SC:adjp
==H:num('um' <card> F S) uma
==A<:pp (das mais antigas discotecas do Algarve)
...
==,
==A<:icl(<pcp>)
===P:vp
====MV:v-pcp('situar' F S) situada
===ADVL:pp
====H:prp('em') em
====P<:np
=====H:prop('Albufeira' F S) Albufeira
==,
...

Coordenação com elementos partilhados

CF8-9 **Se eu dirigisse uma federação**, apresentaria balanços mensais e liberaria minhas contas bancárias.
A1
STA:cu
=CJT:fcl
==ADVL:fcl (Se eu dirigisse uma federação)
===...
==,
==P:vp
===MV:v-fin('apresentar' <nosubj> COND 1S)apresentaria
==ACC:np
===H:n('balanço' <np-idf> M P) balanços
...
=CO:conj-c('e' <co-vfin> <co-fmc>) e
=CJT:fcl
==P:vp
===MV:v-fin('liberar' <nosubj> COND 1S) liberaria
...

Coordenação com elementos partilhados

CF8-9 **Se eu dirigisse uma federação**, apresentaria balanços mensais e liberaria minhas contas bancárias.
A1
STA:fcl
=ADVL:fcl (Se eu dirigisse uma federação)
===...
==,
=X:cu
==CJT:x
===P:vp
====MV:v-fin('apresentar' <nosubj> COND 1S) apresentaria
===ACC:np
====H:n('balanço' <np-idf> M P) balanços
...
==CO:conj-c('e' <co-vfin> <co-fmc>) e
==CJT:x
===P:vp
====MV:v-fin('liberar' <nosubj> COND 1S) liberaria
...

De X a Y

CF26-9 Preço das obras: **de US$ 2.000 a US$ 4.000**.
A1
UTT:np
=H:n('preço' <np-idf> M S) Preço
=N<:pp
==H:prp('de' <sam->) de
==P<:np (as obras)
...
=:
=N<PRED:x
==**PCJT**:pp
===H:prp('de') de
===P<:np
====H:n('US$' <np-idf> M P) US$
====N<:adjp
=====H:num('2.000' <card> M P) 2.000
==**PCJT**:pp
===H:prp('a') a
===P<:np
====H:n('US$' <np-idf> M P) US$
====N<:adjp
=====H:num('4.000' <card> M P) 4.000

Granularidade das etiquetas: *P*, *PMV*, *PAUX*

CP233-5 Podemos beber um copo, jogar gamão, damas, snooker, ver televisão, etc.
A1
STA:fcl
=**PAUX**:v-fin('poder' <nosubj> PR 1P IND) Podemos
=AUX<:cu
==CJT:icl
===**PMV**:v-inf('beber') beber
===ACC:np
====>N:art('um' <arti> M S) um
====H:n('copo' <np-idf> M S) copo
==,
==CJT:icl
===**PMV**:v-inf('jogar') jogar
===ACC:cu
====CJT:np
=====H:n('gamão' <np-idf> M S) gamão
====,

Vagueza estrutural: quantas alternativas?

- **A1, A2…An**
- **F_0/F_1[±x]:f**

CP905-3 Se o IRA apelar a um cessar-fogo **após as eleições** é muito provável que se dê início a um segundo processo de paz.

```
A1
STA:fcl
=ADVL:fcl
==SUB:conj-s('se')          Se
==SUBJ:np (o IRA)
==P:vp
===MV:v-fin('apelar' FUT 3S SUBJ)          apelar
==ADVL:pp
===H:prp('a')     a
===P<:np
====>N:art('um' <arti> M S)      um
====H:n('cessar-fogo' <np-idf> M S)       cessar-fogo
====N</ADVL[-2]:pp
=====H:prp('após')       após
=====P<:np
======>N:art('o' <artd> F P)     as
======H:n('eleição' <np-def> F P)          eleições
```


Floresta Sintá(c)tica vista de dentro

- **2ª fase: reinício em 2007**
- **Refinamento, aperfeiçoamento e incremento**
  - Enfoque em casos em vez de árvores
  - Enfoque em sintagmas nominais
  - Introdução de etiquetas de argumento de nomes (N<ARGS; N<ARGO; N<ARG)
  - Inclusão de corpora de mais géneros que não jornalísticos (ficção, literatura em várias áreas científicas, oral transcrito)

Floresta Sintá(c)tica: 2ªFase

- Melhoria na óptica do **utilizador**:
  - Desenvolvimento do *Milhafre*
  - Lista de procuráveis (estruturas de busca difícil)
  - Melhoria da documentação (Bíblia de referência rápida)

Usos da Floresta

- Treino de sistemas
- Avaliação de sistemas
- Ensino (Freitas 2009)
- Investigação em informática
- **Investigação linguística** …sem grande expressão

A Floresta como recurso para investigação linguística

- Recurso muito rico mas pouco usado
- Razões:
  - Muita informação pode ser contra producente: sistema de etiquetação muito complexo;
  - Percepção (talvez real) de que é preciso conhecer bem o sistema antes de interrogar a Floresta
  - Dificuldade em interagir com os motores de busca
  - Bosque muito pequeno para fazer generalizações

Vantagens em usar a Floresta

- Floresta é um recurso público
- Recurso > uso > melhor recurso > melhor uso > …
- Importância de realizar estudos empíricos (ainda que em combinação com introspecção)
- Testar hipóteses
- Corpora para propósitos diferentes:
  - Bosque tem tamanho suficiente para fazer **estudos qualitativos** (cp. Sampson …e Croft 2009)
  - Selva e Amazónia úteis para **estudos quantitativos**

Estudos quantitativos

- Uso de artigos: com nomes, antes de pronomes possessivos, diferenças entre variantes
- Argumentos internos (que verbos exibem e não exibem argumentos internos, qual a sua forma, dos que não exibem argumentos, quais são transitivos…
- Protótipo de uma categoria

Exemplo de um estudo quantitativo

- Determinação do protótipo de uma categoria (e teste de uma hipótese): as construções de *se* (Afonso, ms)
- Alguma informação de base:
  - Relação entre frequência de estruturas de uso e protótipo
  - Análise colostrutiva (Stefanowitsch andGries2003)
  - Categoria polissémica:

a) Média e b) reflexa

CP455-2 Deixou de ser forçoso as pessoas **deslocarem-se** à sede para **se inscrever** em turnos de 15 dias num dos centros de férias.

c) Recíproca

CP444-2 Nas breves declarações que prestou aos inúmeros jornalistas que **se acotovelavam** à entrada do Palácio…

Construções de *se*

d) Impessoal

CIE-XX-W-1719 **Sabe-se** que Aristóteles interpretou os processos de evaporação e condensação atmosférica…

e) Passiva

FAL-PT-M-587 A gente de ali de fora a ver as portas abrirem-se e a fecharem e as coisas a irem por a porta fora para o rio…

f) Anticausativa

CF109-4 Tal política **convertera-se**, com o avanço da globalização econômica, em sinônimo de proteção ao atraso, ao desperdício e à ineficiência.

Construções de *se*

- Segundo Kemmer (1993), a construção reflexa é o protótipo
- Extracção de todos os verbos que ocorrem com *se*
- Análise colostrutiva: mede o grau de atracção entre verbos e construções
- Agrupamento dos verbos em classes semânticas
- Resultados: os verbos/classes semânticas que ocorrem mais frequentemente nas construções de *se* não indicam reflexividade

words	word freq	obs freq	exp freq	faith	relation	coll.strength
encontrar	13156	6398	39.80	0.4863	attraction	56394
tratar	7424	4843	22.46	0.6523	attraction	46813
tornar	7063	3719	21.37	0.5265	attraction	33513
manter	8927	2696	27.01	0.3020	attraction	20440
deslocar	2178	1483	6.59	0.6809	attraction	14500
mostrar	7884	1996	23.85	0.2532	attraction	14304
queixar	1333	1188	4.03	0.8912	attraction	12878
juntar	3311	1433	10.02	0.4328	attraction	12123
preparar	4510	1426	13.64	0.3162	attraction	10951
limitar	2443	1176	7.39	0.4814	attraction	10279
reunir	4513	1334	13.65	0.2956	attraction	10031
seguir	7704	1462	23.31	0.1898	attraction	9530
fazer	54967	2532	166.30	0.0461	attraction	9210
aproximar	1472	944	4.45	0.6413	attraction	9041
recusar	3067	1064	9.28	0.3469	attraction	8406
referir	8851	1350	26.78	0.1525	attraction	8162
transformar	2108	933	6.38	0.4426	attraction	7944
manifestar	2739	969	8.29	0.3538	attraction	7701
pronunciar	1263	764	3.82	0.6049	attraction	7176
situar	2541	893	7.69	0.3514	attraction	7082

Tabela 1: Verbos mais atraídos pelas construções de *se*

words	word freq	obs freq	exp freq	faith	relation	coll.strength
estar	89398	11	270.46	0.0001	repulsion	449.81
ficar	19829	5	59.99	0.0003	repulsion	85.31
saber	13908	1	42.08	0.0001	repulsion	74.81
garantir	8062	2	24.39	0.0002	repulsion	34.84
receber	8823	5	26.69	0.0006	repulsion	26.69
haver	34793	63	105.26	0.0018	repulsion	19.91
provocar	4412	1	13.35	0.0002	repulsion	19.54
cair	3033	1	9.18	0.0003	repulsion	11.94
assegurar	3429	2	10.37	0.0006	repulsion	10.18
exigir	3680	3	11.13	0.0008	repulsion	8.41
sublinhar	3622	3	10.96	0.0008	repulsion	8.16
acabar	10764	18	32.57	0.0017	repulsion	7.80
ameaçar	2011	1	6.08	0.0005	repulsion	6.56
provar	2011	1	6.08	0.0005	repulsion	6.56
continuar	12779	24	38.66	0.0019	repulsion	6.45
sofrer	3699	4	11.19	0.0011	repulsion	6.16
contribuir	2478	2	7.50	0.0008	repulsion	5.72
acompanhar	3970	5	12.01	0.0013	repulsion	5.27
condenar	2280	2	6.90	0.0009	repulsion	4.85
ceder	1627	1	4.92	0.0006	repulsion	4.66

Tabela 2: Verbos repelidos pelas construções de *se*

Classes de verbos	Exemplos	Frequência
EMOÇÃO	desiludir; chatear; excitar; apaixonar; afeiçoar; orgulhar; arrepender; resignar	98
ALTERAÇÃO FÍSICA	alterar; cozer; evaporar; dilatar; simplificar; aliviar	75
MUDANÇA DE POSSESSÃO	dar; entregar; vender; ocupar; obter; adquirir; coleccionar; conquistar;	74
MOVIMENTO NO ESPAÇO	aproximar/abeirar/achegar; ir; arrastar; subir; sair; correr; fugir; escapar	53
LOCALIZAÇÃO	situar/colocar/pôr/posicionar; centrar; amontoar	53
COGNIÇÃO/VOLIÇÃO/ DESIDERATIVOS	esquecer; lembrar; aperceber; prever; suspeitar; querer; carecer/precisar; esperar	50
CRIAÇÃO	fazer; gerar; transformar; desenhar; fundar; montar	49
COMUNICAÇÃO	dizer; reportar; referir; citar; explicar; propor	47
COMBINAÇÃO/VÍNCULO	misturar; diluir; colar; prender; matricular; vincular	46

Tabela 2: Classes de verbos mais atraídas pelas construções de *se*

Estudos qualitativos

- Anáforas textuais
- Tipos de modificadores de nomes em português
- Posição de advérbios na frase. Que tipos de advérbios existem?
- Categorização de construções: causais, impessoais.

Exemplo de um estudo qualitativo

- Estudo onomasiológico das construções impessoais (Afonso 2008):
  - **Impessoalização** como função comunicativa (supressão/despromoção do agente)
  - Que construções sintácticas veiculam essa função?
  - Corpus oral transcrito (Museu da Pessoa)

Construções impessoais

- Como estudar em corpora estas construções?
  - Procurar ocorrências de construções impessoais obtidas por introspecção ou elicitação
  - Ler o corpus na íntegra e anotar as construções à medida que vão sendo encontradas
  - Existência de um corpus anotado semanticamente e procura de estruturas em que o agente não está marcado

Construções impessoais

- Algumas construções esperadas:
  - construções de *se* (passiva, impessoal e anticausativa)
  - Passiva perifrástica
  - Estratégias lexicais: pronomes pessoais de uso impessoal
- Algumas construções inesperadas:
  - Construções existenciais
  - Nominalizações

Se fosse hoje…

- Aprofundamento da análise das nominalizações, com a introdução de etiquetas de argumento do nome: **N<ARGS/O**
- Acesso à Selva falada, com corpora do PB – comparação de estratégias entre variantes?
- Acesso Selva literária e científica: comparação entre géneros
- Quantificação de estratégias entre variantes e géneros: possibilidade de generalização?

…e o futuro

- Anotação semântica da Floresta: papéis semânticos/do tipo framenet
- Mais revisão:
  - reduzir inconsistências
  - Por casos
  - Mais procuráveis (participação da comunidade precisa-se!)

Será que esta floresta tem futuro?

- Vários outros projectos de florestas para o português têm sido propostos e possivelmente levados a cabo, sem qualquer interacção com o nosso
- O português merece uma floresta de florestas?
- Vale a pena rever um formalismo / um conjunto de frases se não há consenso nem uso nem retorno da comunidade?
- O que é mais importante: a(s) teoria(s) sobre o português, ou um recurso para usar cegamente (Sampson)?

Intermezzo 2

- Apresentação rápida de
  - Floresta Sintá(c)tica: apresentação e história do projecto (parte2)

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('vírus' M S)           vírus
=======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

Perguntas? Opiniões?

- O que gostaria de ver na Floresta?
- O que gostaria de experimentar na Floresta?
- O que é que falta na Floresta?

- Outros analisadores?
- Outras teorias?
- Outra informação?