Na trilha de um teste inter-anotadores

Susana Cavadas Afonso

Projecto: Floresta sintá(c)tica
Última actualização: 8 de Novembro de 2001

O presente texto corresponde ao relatório sobre o procedimento de realização de um teste inter-anotadores, resultados e respectivas conclusões e, mais importante, reflexões para futuros testes de consistência, bem como para a criação ou continuação do processo de criação de um *treebank*.
A formação de categorias foi um processo conjunto com Eckhard Bick com sugestões de Diana Santos. Elaboração dos princípios de contagem e discussão de resultados por Susana Cavadas Afonso e Eckhard Bick.
A forma final do relatório contou com sugestões de Diana Santos.

---

(1) Objectivos:

→ Listagem de situações frequentes de desacordo entre anotadores;
→ Contagem de diferenças observadas entre os ficheiros revistos paralelamente pelos três anotadores;
→ Documentação das diferenças.
→ Contagem das alterações introduzidas no ficheiro automaticamente analisado.

(2) Procedimento:

→ Selecção de 20 extractos contíguos, 107 frases no total, em formato de árvore.
→ Tempo de revisão das frases pelos anotadores: uma semana, cálculo aproximado do tempo que normalmente demoraria a revisão de 100 frases. O tempo limite remete para a igualdade de posição dos anotadores envolvidos no teste, a par das mesmas 100 frases a rever.
→ Revisão intelectual em paralelo das 107 frases directamente em formato de árvores (sem revisão prévia em formato CG) por três anotadores humanos (revisores), sem discussão comum de análise das mesmas durante a fase de revisão dessas ou visualização do resultado da revisão em árvores gráficas.
→ Os anotadores tinham a liberdade de considerarem qualquer tipo de análise, dentro das convenções notacionais, e o número de análises não foi limitado.
→ Comparação dos três ficheiros revistos, dois a dois (**R**(evisão)**1** e **R**(evisão)**2**; **R1** e **R3**; **R2** e **R3**) através do comando de Unix/Linux *diff*.
→ Definição de categorias:
  - Diferenças observadas
  - Origem das diferenças observadas

→ Marcação das diferenças, tendo em conta princípios de contagem (quadro 1);
→ Documentação de situações-problema;
→ Contagem das diferenças observadas e sua origem e apresentação de resultados;
→ Obtenção de uma versão final de análise das 107 frases, após discussão comum entre os anotadores envolvidos no teste sobre a análise linguística mais de acordo

com os critérios notacionais e as opções linguísticas de fundo da Floresta para cada frase.
→ Contabilização do número de análises que na versão final foram provenientes da análise do anotador 1 (A1), anotador 2 (A2), anotador 3 (A3) ou de um acordo/compromisso entre as diferentes análises dos A1, A2, A3.

(3) Descrição das categorias formadas:

3.1. NATUREZA DAS DIFERENÇAS OBSERVADAS:

3.1.1. **Função sintáctica**: diferentes funções atribuídas aos mesmos constituintes entre anotadores.

3.1.2. **Forma sintáctica**:

3.1.2.1 *indentação*: diferenças observadas nos níveis de dependência dos constituintes.

3.1.2.2. ausência *de nó obrigatório / nó extra*: diferenças observadas na ausência/presença de nó não terminal quando a sua criação é obrigatória/ desnecessária porque se está perante um grupo/ uma única folha (nó terminal).

3.1.2.3. *posição do nó*: diferenças observadas na ordem dos constituintes não terminais (nós mãe).

3.1.2.4. *nome de "função" que indica dependência pura*: diferenças observadas em a forma dos nós não terminais (FUNÇÃO: forma), bem como em casos em que o nome da função sintáctica do nó não representa alterações funcionais significativas (cf. Princípios de contagem).

3.1.3. **Morfologia**:

3.3.1.1.género / número
3.3.1.2.classe de palavras
3.3.1.3.lema

3.1.4. **Polilexicais**: anotação de constituintes como unidades separadas (análise em diferentes níveis) ou como uma única unidade (análise em um só nível de constituintes).

3.1.5. **Correcção da separação de frases**

3.1.6. **Etiquetas secundárias**

3.1.7. **Irrelevante**: diferenças observadas não intencionais: espaços, aspas, códigos, meta etiquetas (<sic> / <s frag>, etc.).

3.2. ORIGEM DAS DIFERENÇAS OBSERVADAS:

3.2.1. **Erro humano** (assinalado a castanho nas tabelas de contagem de resultados): análise sintáctica incorrecta.
As diferenças observadas teriam como origem erro humano numa situação como a seguinte:
Exemplo: Isto é um exemplo!

ANOTADOR 1	ANOTADOR 2
EXC:fcl	**STA**:fcl

SUBJ:pron-indp('isto' <dem> M S) Isto P:v-fin('ser' PR 3S IND) é **SC**:np =>N:art('um' <arti> M S) um =H:n('exemplo' M S) exemplo !	**SC**:pron-indp('isto' <dem> M S) Isto P:v-fin('ser' PR 3S IND) é **SUBJ**:np =>N:art('um' <arti> M S) um =H:n('exemplo' M S) exemplo !

3.2.2. **Existência de diferentes análises aceites** (a verde): diferentes análises entre anotadores, aceites por todos os anotadores como análises válidas. As diferentes análises aceites encontram-se documentadas.

Exemplo: A Judiciária aproveitou ainda o balanço para passar buscas à casa de Reinaldo Teles.

ANOTADOR 1	ANOTADOR 2
A1 STA:fcl SUBJ:np =>N:art('a' F S) A =H:prop('Judiciária' F S) Judiciária P:v-fin('aproveitar' PS 3S IND) aproveitou **ADVL**:adv('ainda') ainda ACC:np =>N:art('o' M S) o =H:n('balanço' M S) balanço ADVL:pp =H:prp('para') para =P<:icl ==P:v-inf('passar') passar ==ACC:n('busca' F P) buscas ==ADVL:pp ===H:prp('a' <sam->) a ===P<:np ====>N:art('a' <-sam> F P) as ====H:n('casa' F P) casas ====N<:pp =====H:prp('de') de =====P<:prop('Reinaldo_Teles' M S) Reinaldo_Teles	A1 STA:fcl SUBJ:np =>N:art('a' F S) A =H:prop('Judiciária' F S) Judiciária P:v-fin('aproveitar' PS 3S IND) aproveitou ACC:np **=>N**:adv('ainda') ainda =>N:art('o' M S) o =H:n('balanço' M S) balanço ADVL:pp =H:prp('para') para =P<:icl ==P:v-inf('passar') passar ==ACC:n('busca' F P) buscas ==ADVL:pp ===H:prp('a' <sam->) a ===P<:np ====>N:art('a' <-sam> F P) as ====H:n('casa' F P) casas ====N<:pp =====H:prp('de') de =====P<:prop('Reinaldo_Teles' M S) Reinaldo_Teles

3.2.3. **Existência de diferentes análises rejeitadas** (a vermelho): categoria distinta de erro humano. Os anotadores apresentaram análises distintas e uma das análises, apesar de poder ser aceite em termos puramente sintácticos, foi rejeitada, por não cumprir as opções linguísticas tomadas (Ex: caso de o sinal de pontuação conter valor

sintáctico- exemplo abaixo- obrigando a uma determinada estrutura sintáctica).

Exemplo: A maçã está podre, esteve muito tempo ao sol.

ANOTADOR 1	ANOTADOR 2
A1 **STA:cu** **CJT:fcl** =SUBJ:np ==>N:art('a' <artd> F S) A ==H:n('maçã' F S) maçã =P:v-fin('estar' PR 3S IND) está =SC:adj('podre' F S) podre , **CJT:fcl** =P:v-fin('estar' PS 3S IND) esteve =ADVL:n('muito_tempo' M S) muito_tempo =ADVS:pp ==H:prp('a' <sam->) a ==P<:np ===>N:art('o' <-sam> M S) o ===H:n('sol' M S) sol .	A1 STA:fcl SUBJ:np =>N:art('a' <artd> F S) A =H:n('maçã' F S) maçã P:v-fin('estar' PR 3S IND) está SC:adj('podre' F S) podre , ADVL:fcl =P:v-fin('estar' PS 3S IND) esteve =ADVL:n('muito_tempo' M S) muito_tempo =ADVS:pp ==H:prp('a' <sam->) a ==P<:np ===>N:art('o' <-sam> M S) o ===H:n('sol' M S) sol .

3.2.4. **Outras análises** (a roxo): entre anotadores, existe pelo menos uma análise em comum e pelo menos uma análise distinta e que é aceite:

- análise(s) extra expressa(s) por uma análise completa (A1 e A2, sendo A1 ou A2 comum entre os anotadores). Estes casos foram contabilizados em função sintáctica.
- análise(s) extra expressa(s) na mesma linha da frase, através do formalismo de representação de ambiguidade (cf. www.visl.hum.sdu.dk/pt/guidelines.html) tanto a nível de forma como de função sintáctica. Exemplos:
  - Anotador 1: N<PRED:pp
    Anotador 2: N<PRED/ADVL[-1]:pp
  - Anotador 1: N<PRED: pp
    Anotador 2: N<PRED:pp/acl

Exemplo: Estavam repeltas de lixo, copos de plástico sujos de café.

ANOTADOR 1	ANOTADOR 2
A1 STA:fcl P:v-fin('estar' IMPF 3P IND) Estavam SC:ap	A1 STA:fcl P:v-fin('estar' IMPF 3P IND) Estavam SC:ap

=H:adj(’repleto’ F P) repletas =A<:pp ==H:prp(’de’) de ==P<:cu ===CJT:n(’lixo’ M S) lixo ===, ===CJT:np ====H:n(’copo’ M P) copos ====N<:pp =====H:prp(’de’) de =====P<:np ======H:n('plástico' M S) plástico ======N<:ap =======H:adj('sujo' M P) sujos =======A<:pp ========H:prp('de') de ========P<:n('café' M S) café . **&&** **A2** **STA:fcl** **P:v-fin('estar' IMPF 3P IND) Estavam** **SC:ap** **=H:adj('repleto' F P) repletas** **=A<:pp** **==H:prp('de') de** **==P<:np** **===H:n('lixo' M S) lixo** **===,** **===N<PRED:np** **====H:n('copo' M P) copos** **====N<:pp** **=====H:prp('de') de** **=====P<:n('plástico' M S) plástico** **====N<:ap** **=====H:adj('sujo' M P) sujos** **=====A<:pp** **======H:prp('de') de** **======P<:n('café' M S) café** **.**	=H:adj('repleto' F P) repletas =A<:pp ==H:prp('de') de ==P<:cu ===CJT:n('lixo' M S) lixo ===, ===CJT:np ====H:n('copo' M P) copos ====N<:pp =====H:prp('de') de =====P<:np ======H:n('plástico' M S) plástico ======N<:ap =======H:adj('sujo' M P) sujos =======A<:pp ========H:prp('de') de ========P<:n('café' M S) café .

3.2.5. **Variante do português** (a azul): diferentes análises derivadas das diferenças estruturais entre o português europeu e o português do Brasil, por exemplo, meses do ano cuja grafia em português do Brasil e em português europeu diverge (letra minúscula ou maiúscula, respectivamente)

3.2.6. **Conhecimento extra-linguístico dos diferentes anotadores** (a rosa): exemplo, género/ número de nomes próprios não visíveis pelo contexto.

(4) Resultados:

I. Considerações gerais relativamente aos resultados e contagem das diferenças observadas:

- Princípios de contagem: (ver quadro 1 com os respectivos exemplos):

a) *Princípio da exclusividade*: apenas uma diferença (função sintáctica) é contabilizada se as diferenças observadas residirem em funções sintácticas, que são exclusivas, observadas no mesmo nível de constituintes

Anotador 1:	F1:x	y
	F2:x’	y’
Anotador 2:	F2:x	y
	F1:x’	y’

F1 e F2 são funções exclusivas no mesmo nível de constituintes. Duas diferenças observadas, mas uma diferença contabilizada, uma vez que a alteração de uma das etiquetas sintácticas produziria automaticamente uma mudança na outra etiqueta em questão.

b) *Prevalecimento da função sintáctica sobre a forma sintáctica*: a dependência da forma sintáctica em relação à função sintáctica, isto é, determinação da forma sintáctica pela função sintáctica, implica a contagem de uma diferença em função sintáctica.

Anotador 1:	F1:x	y
	[i]F2:x’	y’
Anotador 2:	F1:x	y
	[i+z] F3:x’	y’

sendo [i], indentação (forma sintáctica) e perante a condição

se A1 (F2) → A2 (F3), logo A1 ([i]) → A2 ([i+z]),

apenas uma diferença em função sintáctica é contabilizada.

c) Polilexicais: o uso diferenciado de uma análise em vários níveis ou num único nível de um sintagma nominal, produz diferenças a nível sintáctico **internas** à estrutura do polilexical que são inteiramente dependentes da forma sintáctica adoptada, daí, a contagem de uma diferença em

polilexicais. No entanto, ao nível mais alto, oracional (externas à própria estrutura do polilexical), se se observarem diferenças na função sintáctica entre o polilexical e o seu correspondente em vários níveis, essa diferença será contabilizada de forma regular.

d) *Função sintáctica dependente da forma sintáctica:*

c2) Nó não terminal (nó mãe) desnecessário: a criação desnecessária de um grupo acarreta diferenças na forma sintáctica, nomeadamente na indentação (descida de pelo menos um nível constituinte). Nestes casos, apenas uma diferença em ausência de nó obrigatório / nó extra é contabilizada, por as diferenças de indentação serem exclusivamente dependentes da criação do nó.

c3) Constituintes partilhados: Em relações de coordenação, a partilha de constituintes pelas duas orações coordenadas versus a não partilha dos constituintes (isto é, incluidos apenas em uma das orações coordenadas) é contabilizada como uma diferença em forma sintáctica (forma), independentemente das diferenças em função sintáctica, e indentação que daí decorrem.

e) *Etiquetas de dependência de constituintes*: As etiquetas N< e A<, tal como N<PRED e N< apesar de corresponderem a função, apenas indicam a que nível o constituinte está posicionado. Uma diferença em forma sintáctica (*nome de "função" que indica dependência pura*) é contabilizada.

Resumo e exemplos dos princípios de contagem adoptados aplicados às categorias de diferenças observadas, no quadro 1.

- Sistema de contagem:

a) As diferenças foram contabilizadas unidireccionalmente, isto é, contabilizou-se o número de diferenças entre R1 e R2, mas não entre R2 e R1 distintamente, ou seja, foi apenas contabilizado uma vez as diferenças observadas entre o revisor 1 (R1) e o revisor 2 (R2).
b) cada diferença observada entre R1 e R2, R1 e R3; R2 e R3 foi contabilizada, independentemente do facto de ser a mesma diferença observada entre R1 e R2 ou R1 e R3:

Caso 1: 3 análises distintas
R1: análise A
R2: análise B
R3: análise C

R1 diff. R2 : A vs. B → 1 diferença contabilizada
R1 diff. R3: A vs C → 1 diferença contabilizada
R2 diff. R3: B vs. C → 1 diferença contabilizada

3 diferenças contabilizadas no total

Caso 2:

R1: análise A
R2: análise A
R3: análise B

R1 diff. R2: A vs A → 0 diferenças contabilizadas
R1 diff. R3: **A vs.B** → 1 diferença contabilizada
R2 diff. R3: **A vs.B** → 1 diferença contabilizada } 2 dif., mesma situação (A vs.B)

2 diferenças contabilizadas

c) O cálculo dos totais de diferenças, correspondem à soma de R1 *diff.* R2; R1 *diff.* R3; R2 *diff.* R3;
d) As categorias foram definidas antes do processo de comparação e ao longo deste, as diferenças observadas iam sendo classificadas segundo essas categorias;
e) Não foram contabilizadas R1, R2 e R3 *diff.* Af (análise final);
f) Os erros humanos não foram subcategorizados, ou seja, não foi contabilizada a razão que levou à produção de diferentes análises por parte dos diferentes anotadores: por exemplo, interpretação, falta de atenção, entre outras possíveis. Por exemplo, especificamente em relação à função sintáctica, foram contabilizadas, como diferença sintáctica, diferenças no nó ao mais alto nível STA/QUE/COM/UTT/(…) cuja origem é erro humano (que sem subcategorização poderia ser associado a falha/ erro linguística, o que, de facto, não corresponde, neste caso à realidade). Em relação a morfologia, a causa da diferença contabilizada resultante da não desambiguização foi também atribuída a erro humano.

g) Diferenças não contabilizadas, por serem derivadas de:

f1) edição de texto:
- formalismo (ausência de dois pontos entre FUNÇÃO: forma);
- Coexistência de funções exclusivas entre si (FUNÇÃO1:FUNÇÃO2: forma);
- Espaços;

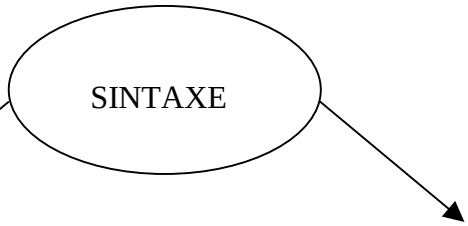
f2) apagamento de linhas da árvore
f3) formalismo CG, como o uso das setas de dependência em formato de árvore (FUNÇÃO>:forma).

- Apresentação dos resultados:

Apresentam-se os resultados obtidos nas categorias Sintaxe (função e forma), Morfologia (classe de palavras, género e número e lema), polilexicais e correcção da separação de frases.

Dos resultados obtidos nas restantes categorias, nomeadamente as categorias *Irrelevante* ou *etiquetas secundárias*, as ilações não parecem ser produtivas.

A) Contabilização de diferenças observadas em Sintaxe
*(percentagens referentes ao **número total de diferenças** contabilizadas, não às 107 frases revistas)*



FORMA (total: 360)	FUNÇÃO
o *Indentação*: 133 diferenças observadas em 51 frases; 0 diferenças observadas em 56 frases.	o 303 diferenças observadas em 74 frases 0 diferenças observadas em 33 frases
o *Ausência de nó obrigatório/ nó extra*: 116 diferenças em 41 frases; 0 diferenças observadas em 66 frases	
o Posição do nó: 27 diferenças observadas em 15 frases; 0 diferenças observadas em 92 frases. o *Forma*: 84 diferenças observadas em 32 frases; 0 diferenças observadas em 75 frases.	
Erro humano: 95,4 % ; 80,4% ; 62,9% ; 82,1%	Erro humano: 77,8%
Diferentes análises aceites: 3% ; 6,9% ; 37% ; 17,8%	Diferentes análises aceites: 15,8%
Diferentes análises não aceites: 0%	Diferentes análises não aceites: 0,66%
Outras análises (uma em comum): 1,5% ; 0,9%	Outras análises (uma em comum): 5,61%
Variante do português: 0%	Variante do português: 0%
Conhecimento do mundo: 0%	Conhecimento do mundo: 0%

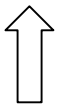
B) Contabilização das diferenças observadas em Morfologia
*(percentagens referentes ao **número total de diferenças** contabilizadas, não às 107 frases revistas)*

Classe de palavras	Lema	Género/número
15 diferenças observadas em 12 frases; 0 diferenças observadas em 95 frases	3 diferenças observadas em 2 frases; 0 diferenças observadas em 105 frases	74 diferenças obsrervadas em 72 frases; 0 diferenças observadas em 35 frases
		
Erro humano: 100%	Erro humano: 33,3%	Erro humano: 94,5%
	Variante do português: 66,6%	Conhecimento do mundo: 5,4%

(Nota: As outras categorias de origem da diferenças observadas que não indicadas na tabela acima tiveram 0% de ocorrências)

C) Contabilização das diferenças observadas em Polilexicais
*(percentagens referentes ao **número total de diferenças** contabilizadas, não às 107 frases revistas)*

14 diferenças observadas em 8 frases ; 0 diferenças observadas em 99 frases



- Erro humano: 21,4%
- Diferentes análises aceites: 78,5%

(Nota: As outras categorias de origem da diferenças observadas que não indicadas na tabela acima tiveram 0% de ocorrências)

D) Contabilização das diferenças observadas em Correcção da separação de frases
*(percentagens referentes ao **número total de diferenças** contabilizadas, não às 107 frases revistas)*

4 diferenças observadas em 2 frases ⇦ 100% erro humano

- Interpretação dos resultados:

Em valores totais, observaram-se 785 diferenças em 107 frases nas várias categorias formadas. Este valor não inclui os valores das diferenças obtidas em diferenças irrelevantes.

Em termos de diferenças observadas por frase, a média foi de 7,3 diferenças por frase, não contando com as diferenças irrelevantes. Porém, observe-se que uma única frase exibiu, no total, 106 diferenças.

Tendo em conta os valores obtidos em cada uma das categorias, há a considerar que:

1) a nível de função sintáctica, as diferenças observadas entre os anotadores teve como origem principal erro humano (em 77,8% dos casos). Não se especificou o que poderia ter levado ao erro humano: diferença de paradigma, falta de atenção, erro (literalmente), convenções pré-definidas e não implementadas (ex: STA ou UTT, no nó mais alto).
   Em termos de média de diferenças por frase a este nível, estimam-se 4,09 diferenças por frase (303 diferenças em 74 frases) com um índice de 2,2 diferenças por erro humano por frase. De novo, a chamada de atenção para o facto de a frase C48-4 exibir isoladamente 41 diferenças em função sintáctica, isto é, 55,4% do total de diferenças nesta categoria.

2) Apesar de comparativamente ao valor das diferenças obtidas na função sintáctica, os valores nas categorias *Indentação* e *Ausência de nó obrigatório/ nó extra* não serem relevantes, a ausência de nó obrigatório/ nó extra foi um aspecto recorrente neste teste, a ser levado em conta pelos anotadores nos próximos testes e no processo de revisão em geral.
   O número elevado de erros humanos verificado a nível da *forma sintáctica* nas árvores que foram o objecto de escrutínio no presente teste pode explicar--se parcialmente pelo facto de a ferramenta de edição de árvores estar a ser desenvolvida em paralelo e ter, por sua vez, nas suas várias versões, introduzido problemas inesperados. Espera-se que futuros testes já possam ser realizados sem esta desvantagem.

Para o elevado número de diferenças cuja causa foi erro humano (646, no total, correspondendo a uma média de 6,03 erros humanos por frase nas várias categorias) pode ter contribuído:

a) **a complexidade e tamanho das frases**: inspeccionou-se individualmente as frases que exibiam o maior número de diferenças observadas e constatou-se que eram:

   - muito longas (42 palavras por frase em média, num intervalo entre [8-88]);
   - complexas (mais de uma oração subordinada ou coordenada, muitos níveis de constituintes- tornando difícil por vezes a tarefa de análise/revisão do nível dos constituintes-, envolvendo descontinuidades, discurso indirecto livre);
   - iniciadas por uma preposição, conjunção integrante, pronomes relativos ou frases sem verbo principal;
   - listas;

b) **a alteração no processo de revisão:**

   d1) a não visualização das árvores invertidas, por questões técnicas: os anotadores, ao contrário de todo o processo de revisão anterior, não visualizaram as árvores gráficas, que auxilia na detecção de problemas principalmente a nível de indentação e ausência de nós obrigatórios ou nós extra, porque os ficheiros de um anotador seriam constantemente alterados pelos ficheiros dos outros anotadores, uma vez que se tratava do mesmo conjunto de frases a ser visualizado.

   d2) a revisão directa em formato de árvores por questões de consumo de tempo: a revisão foi feita exclusivamente em formato de árvores, ou seja,

não houve uma fase de revisão do formato CG. O que significa que o anotador não teve a oportunidade de previamente focar o processo de revisão na informação sintáctica e morfológica. Os ficheiros em formato de árvore não foram portanto gerados a partir dos ficheiros CG e o anotador não pôde concentrar a sua atenção apenas no nível de constituintes e nós. Talvez este facto explique também o elevado número de erro humano em termos de morfologia género/número (basicamente, não desambiguização).

Conclusões :

- Face aos objectivos inicialmente propostos, não foram medidas as diferenças/alterações introduzidas no ficheiro gerado automaticamente.
- A maioria das diferenças observadas situam-se a nível da forma e função sintáctica, sendo o erro humano o principal motivo para a ocorrência dessas diferenças. A ausência de nó obrigatório na formação de grupos sob a categoria de forma sintáctica, foi um aspecto recorrente a ter em conta em futuros testes.
  Apesar das diferenças observadas em função sintáctica serem maioritariamente derivadas de erro humano, verificou-se, no entanto, durante a fase da discussão de análises, um elevado grau de concordância relativamente à versão final. Este facto, tendo igualmente em conta que a versão final foi em 33% (contra 20,3%, versão anotador A, 16,9%, anotador B e 6,7%, anotador C) um acordo entre os três anotadores, pode ser indicador de que, de modo a conseguir a consistência desejada em termos de análise, as mesmas frases teriam de ser revistas por, no mínimo, dois anotadores.
- Erro humano justifica integralmente diferenças observadas a nível de morfologia, género/número. A interpretação deste fenómeno está ligada à não revisão de desambiguização por parte dos anotadores, do que se conclui que os anotadores focaram a sua atenção na revisão da sintaxe. Este é um aspecto a ter em mente em futuras revisões, pois é um problema de fácil colmatação.

Direcções a seguir:

- QUANTO AO TESTE INTER-ANOTADORES

  I. De forma a confirmar a hipótese levantada de que a revisão consecutiva, não paralela, das mesmas frases por dois ou mais anotadores poderia eventualmente reduzir o número de diferenças observadas, um conjunto de frases, em número inferior a 100, poderia ser revisto de forma faseada: o anotador 1 revê relativamente a parâmetros específicos, e limitados, a análise automática, anotando a ocorrência de alterações introduzidas; o anotador 2, revê igualmente, dentro dos mesmos parâmetros, a análise automática do analisador sintáctico e revê as alterações introduzidas pelo anotador 1. Se o resultado da precisão for elevado, a hipótese está confirmada.

  II. É desaconselhável a tentativa de medição de várias questões num único processo: teste inter-anotadores. É de conhecimento geral que há questões de mais fácil avaliação do que

outras (Brants, 2000). Deste modo, será uma medida de avaliação mais concreta e real, a realização de testes mais focados em questões mais restritas.

IV. Na formação da categoria *Diferentes análises aceites ou Outras análises* (Origem da observação das diferenças), seria interessantes, em futuros testes, avaliar a que nível / níveis essas análises foram aceites: diferentes interpretações ou diferenças na notação utilizada. Um exemplo concreto neste teste é o exemplo incluido no ponto 3.2.4. (Estavam repletas de lixo, copos de plásticos sujos de café), em que a aceitabilidade de uma outra análise não comum entre os anotadores foi ao nível da interpretação (que por sua vez desencadeou diferenças igualmente a nível notacional).

V. Racionalização do processo de comparação:

- → impondo condições/limitações ao comando *diff*: por exemplo, comparação entre os três ficheiros, ignorando espaços, etc.
- → Focar a comparação em questões que à partida conduzam a conclusões, tendo em conta, sempre, os objectivos inicialmente propostos.

VI. Racionalização da contagem: de forma a controlar/verificar a contagem das diferenças entre os anotadores, seria talvez útil, numa fase de planeamento do teste, prever que tipo de situações poderiam ocorrer. Por exemplo, entre três ficheiros, seria impossível a obtenção de uma diferença absoluta por frase:

se e = r ^ r = s => e = s (zero diferenças no total)

VII. A contabilização das diferenças introduzidas por cada anotador, ou pela versão final, em relação à versão automática original não foi realizada, mas seria interessante fazê-lo e tentar medir o nível de ruído do analisador sintáctico e do analisador/revisor humano.

- QUE FUTURO?

I. Em termos de anotação de um corpus (revisão manual/intelectual de um corpus pré-anotado), uma das formas possivelmente mais equilibrada de o fazer seria a divisão deste em partes com diferentes níveis de especificação e também de perfeição (percentagem de erros), ou seja, uma das partes (10%), seria sujeita a uma revisão exaustiva, a todos os níveis; outra(s) parte(s) (até 50%) seria revista 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 uma discussão de casos-problema, amplificação do léxico, adaptação do analisador sintáctico relativamente ao corpus em questão. Um exemplo real retirado da Floresta Sintá(c)tica é a questão das datas. Primeira fase, correspondentes a 10 % do corpus acima mencionado: eliminação da categoria neutra / não especificada e discussão do número /género de datas. Segunda fase, sistematização e automatização da decisão acordada (M S), ou seja, adaptação e melhoramento do analisador sintáctico e, consequentemente, anotação automática do restante corpus, que exibe a característica mencionada.

- Em estreita ligação com o ponto anterior, note-se que, lidando com um corpus de grandes dimensões, um compromisso entre qualidade/ambição e quantidade/realidade tem de ser levado em conta. No caso específico da Floresta Sintá(c)tica, criaram-se mais categorias e distinções mais finas, o que resultou em uma quantidade mais elevada de erros, porque (1) o analisador sintáctico não foi optimizado de acordo com as novas categorias, (2) as distinções mais finas exigiram mais intervenção humana além de levantarem outras questões dúbias relacionadas e (3) essas distinções interferem com outras categorias existentes. Por exemplo, de novo a tentativa de eliminação da etiqueta neutra / não especificada F/M, relativamente a nomes topológicos (cidades, países e afins). O problema directamente relacionado com esta tentativa de distinção mais fina foi os topónimos que não requerem artigo, como Aveiro ou Campinas, implicando maior intervenção humana e menor automatização, o que se resume em uma maior qualidade (maior especificação) mas em uma menor quantidade e consistência.

- De modo a reduzir o número de diferenças por ”erro humano”, o desenvolvimento de ferramentas de auxílio à revisão que impedissem, marcassem para inspecção posterior ou corrigissem certas operações porque incompatíveis em termos formais, seria muito útil.

- Tendo em conta o elevado número de diferenças observadas por erro humano, há que equacionar de que forma se poderá, desde cedo, promover a maior eficiência dos anotadores: uma das formas poderia ser o treino de anotadores numa fase de pré-construção de um treebank. Pelos resultados em termos de listagem de situações recorrentes que surgiram durante o presente teste das quais os anotadores não estavam plenamente conscientes, outra forma seria a realização mais frequente de testes deste género (de forma mais restrita em termos do objecto de avaliação), para identificação de problemas mais prematuramente e daí definir estratégias para a sua colmatação. À partida parece que as discussões regulares debruçaram-se essencialmente sobre questões de função sintáctica ou de estrutura (que tipo de solução a encontrar para problemas de difícil resolução em termos linguísticos), e questões de base como a formação de grupos (exactamente um dos problemas recorrentes que emergiu da comparação das análises entre os três anotadores) tornaram-se transparentes.

Quadro 1: Distribuição de princípios de contagem por categorias de diferenças observadas.

Diferenças observadas		Forma sintáctica				Morfologia			Género de nomes próprios	Polilexicais	Correcção de separação frásica	Etiquetas secundárias	Irrelevante
Princípios orientadores da contagem	Função sintáctica	Indentação	Ausência de nó obrigatório/ nó extra	Posição de nó errada	Nome da "função" que indica dependência pura	Classe de palavras	Lema	Género/ número					
Contagem geral, independente (a)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Prevalecimento da função sintáctica sobre a forma sintáctica (b)	✓	✓	✓	✓	✓								
Exclusividade sintáctica (c)	✓												
Diferenças dependentes da forma sintáctica (d)	✓		✓		✓					✓			

(a) uma diferença observada correspondendo a uma diferença contabilizada

(b) mais do que uma diferença observada correspondendo a uma diferença contabilizada

Ex: SOURCE: CETEMPúblico n=48 sec=clt sem=97a

C48-8 Depois, poderá haver explorações em etapas sucessivas para outras linguagens.

Análise 1	Análise 2
A1 STA:fcl ADVL:adv('depois') Depois , P:vp =AUX:v-fin('poder' FUT 3S IND) poderá =MV:v-inf('haver') haver ACC:n('exploração' F P) explorações ADVL:pp =H:prp('em') em =P<:np ==H:n('etapa' F P) etapas ==N<:adj('sucessivo' F P) sucessivas **==N<:pp** ===H:prp('para') para ===P<:np ====>N:pron-det('outro' <diff> F P) outras ====H:n('linguagem' F P) linguagens	A1 STA:fcl ADVL:adv('depois') Depois , P:vp =AUX:v-fin('poder' FUT 3S IND) poderá =MV:v-inf('haver') haver ACC:n('exploração' F P) explorações ADVL:pp =H:prp('em') em =P<:np ==H:n('etapa' F P) etapas ==N<:adj('sucessivo' F P) sucessivas **ADVL:pp** =H:prp('para') para =P<:np ==>N:pron-det('outro' <diff> F P) outras ==H:n('linguagem' F P) linguagens

Diferenças observadas:

a) 1 diferença em função sintáctica;
b) 1 forma sintáctica (indentação), dependente de a)

Diferenças contabilizadas: 1 diferença em função sintáctica

(c) duas diferenças observadas a nível da função sintáctica correspondendo a uma diferença contabilizada

EX: C46-5 Essa é uma pergunta que ainda hoje permanece sem resposta

Análise 1		Análise 2	
STA:fcl		STA:fcl	
SUBJ:pron-det('esse' <dem> F S)	Essa	SC:pron-det('esse' <dem> F S)	Essa
P:v-fin('ser' PR 3S IND)	é	P:v-fin('ser' PR 3S IND)	é
SC:np		SUBJ:np	
=>N:art('um' <arti> F S)	uma	=>N:art('um' <arti> F S)	uma
=H:n('pergunta' F S)	pergunta	=H:n('pergunta' F S)	pergunta
=N<:fcl		=N<:fcl	
==SUBJ:pron-indp('que' <rel> F S)	que	==SUBJ:pron-indp('que' <rel> F S)	que
==ADVL:advp		==ADVL:advp	
===>A:adv('ainda')	ainda	===>A:adv('ainda')	ainda

===H:adv('hoje') hoje ==P:v-fin('permanecer' PR 3S IND) permanece ==SC:pp ===H:prp('sem') sem ===P<:n('resposta' F S) resposta	===H:adv('hoje') hoje ==P:v-fin('permanecer' PR 3S IND) permanece ==SC:pp ===H:prp('sem') sem ===P<:n('resposta' F S) resposta

Diferenças observadas (2 dif.)

1. SUBJ:pron-det *vs.* SC:np
2. SC:np *vs.* SUBJ:np

Diferenças contabilizadas (1 dif.)

1 diferença contabilizada: pelo princípio da exclusividade, a revisão de uma das etiquetas (SUBJ ou SC) produziria a alteração imediata da outra etiqueta envolvida (SC ou SUBJ), isto é, não seria provável a existência a mesmo nível de dois SUBJ ou de dois SC

(d) mais do que uma diferença observada radicadas na forma sintáctica correspondendo a uma diferença contabilizada

Ex: (…) o atentado poderia ter partido de a velha guarda do clã Gambino.

Análise 1	Análise 2
A1 STA:fcl (…) SUBJ:np =>N:art('o' <artd> M S) O =H:n('atentado' M S) atentado P:vp =AUX:v-fin('poder' COND 3S) poderia =AUX:v-inf('ter') ter =MV:v-pcp('partir') partido ADVL:pp =H:prp('de' <sam->) de =P<:np ==>N:art('a' <-sam> F S) a ==>N:adj('velho' F S) velha ==H:n('guarda' F S) guarda ==N<:pp ===H:prp('de' <sam->) de ===P<:np ====>N:art('o' <-sam> M S) o ====H:n('clã' M S) clã ====N<:prop('Gambino' M S) Gambino .	A1 STA:fcl (…) SUBJ:np =>N:art('o' <artd> M S) O =H:n('atentado' M S) atentado P:vp =AUX:v-fin('poder' COND 3S) poderia =AUX:v-inf('ter') ter =MV:v-pcp('partir') partido ADVL:pp =H:prp('de' <sam->) de =P<:np ==>N:art('a' <-sam> F S) a ==H:np('velha_guarda' F S) velha_guarda ==N<:pp ===H:prp('de' <sam->) de ===P<:np ====>N:art('o' <-sam> M S) o ====H:n('clã' M S) clã ====N<:prop('Gambino' M S) Gambino .

Diferenças observadas:

c) forma sintáctica: indentação, forma
d) função sintáctica

Diferenças contabilizadas: 1 diferença em ”polilexicais”

(NOTA: no entanto, se se observar diferenças na função sintáctica do polilexical na oração principal, isto é, se as diferenças entre o uso da forma polilexical e análise das suas partes constituintes não se resumirem à estrutura interna do próprio polilexical, então, a diferença deverá ser contabilizada. Por exemplo, imaginemos a situação: H:np *versus* >N:np. Neste caso, a diferença observada não deriva das alterações à estrutura interna do polilexical, logo, uma diferença deveria ser também contabilizada a nível da função sintáctica)