Ensaio sobre a revisão da oralidade

Paula Taveira & Diana Santos

anagoncil@gmail.com, d.s.m.santos@ilos.uio.no

Encontro da APL, 28 de outubro de 2015

Em duas palavras

Rever as entrevistas portuguesas do Museu da Pessoa incluídas no AC/DC

1. Enquadramento
2. Questões metodológicas
3. Resultados
4. Considerações

1. Enquadramento

- A Linguateca pós-financiamento
- O AC/DC e a Gramateca
- O Museu da Pessoa e o corpo Museu da Pessoa
- A nossa opinião sobre a língua portuguesa

A Linguateca pós-financiamento

- A partir de 2010 não foi possível financiar a Linguateca por causa das regras europeias que entraram em vigor em Portugal
- A partir daí, quase toda a atividade da Linguateca é feita
  - em regime gratuito/voluntário
  - com financiamentos pequenos, pontuais, de diferentes entidades

O que hoje aqui apresentamos é resultado de trabalho voluntário das duas autoras, e do apresentador, e o seu fim é continuar a enriquecer os recursos servidos pela Linguateca.

O AC/DC

O AC/DC é o projeto mais antigo da Linguateca: Acesso a Corpos / Disponibilização de Corpos

- Existe desde 1999
- Serve à comunidade interessada na linguística do português muitos corpos diferentes, todos eles com uma anotação sintática efetuada pelo PALAVRAS (Bick, 2000) e com anotação semântica, cf. Santos (2014)
- Tem associados vários serviços de procura e comparação (Simões & Santos, 2014)

Um dos corpos incluídos no AC/DC é o Corpo Museu da Pessoa.

Usos do Corpo Museu da Pessoa

- Para estudos de linguística em geral
- Para o ensino de português como língua estrangeira
- Para a Gramateca (novo projeto no âmbito da Linguateca para fazer gramática baseada em corpos), `http://www.linguateca.pt/Gramateca/`

O Museu da Pessoa: porquê a sua revisão?

O Museu da Pessoa é um museu virtual e colaborativo fundado em São Paulo no ano de 1991. Desde sua origem, tem como objetivo registrar, preservar e transformar em informação, histórias de vida de toda e qualquer pessoa da sociedade. Nosso acervo conta atualmente com mais de 16 mil depoimentos em áudio, vídeo e texto e cerca de 72 mil fotos e documentos digitalizados. `http://www.museudapessoa.net/pt/entenda/o-museu-da-pessoa`

- O Núcleo Português do Museu da Pessoa estava sedeado na Universidade do Minho e cedeu-nos inicialmente cem entrevistas digitalizadas por alunos estagiários (ver Almeida et al., 2000)
- Mas: essas páginas continham bastantes erros de ortografia (e provavelmente de transcrição também). Como era dos poucos casos em que o AC/DC tinha linguagem oral, pareceu-nos importante melhorar esse recurso.

A situação

No ciclo de tratamento de uma história de vida... (de Almeida et al., 2000)

1. reprodução fiel do discurso registado: as pausas, os risos, as repetições, (...) e as variações morfológicas devem ser cuidadosamente anotadas
2. revisão da entrevista transcrita... entrevista editada, reagrupada por temas...
3. indexação

Pensamos que obtivemos as entrevistas após o passo 1, embora faltassem as indicações de risos, suspiros, assobios, etc., possivelmente retiradas pelo nosso próprio processamento do XML para texto.

A tarefa a que nos propusemos

- Rever a língua, mantendo as marcas da oralidade, mas corrigindo os erros dos transcritores e dos falantes – que, como se verá, eram muitos.

Nesse processo, acabámos por marcar/anotar os erros ou desvios relativos à norma padrão portuguesa conhecida e praticada pela primeira autora, revisora de profissão, após discussão dos casos mais complexos pelas duas autoras, e sua documentação em `http://www.linguateca.pt/acesso/revisao_mp.html`.

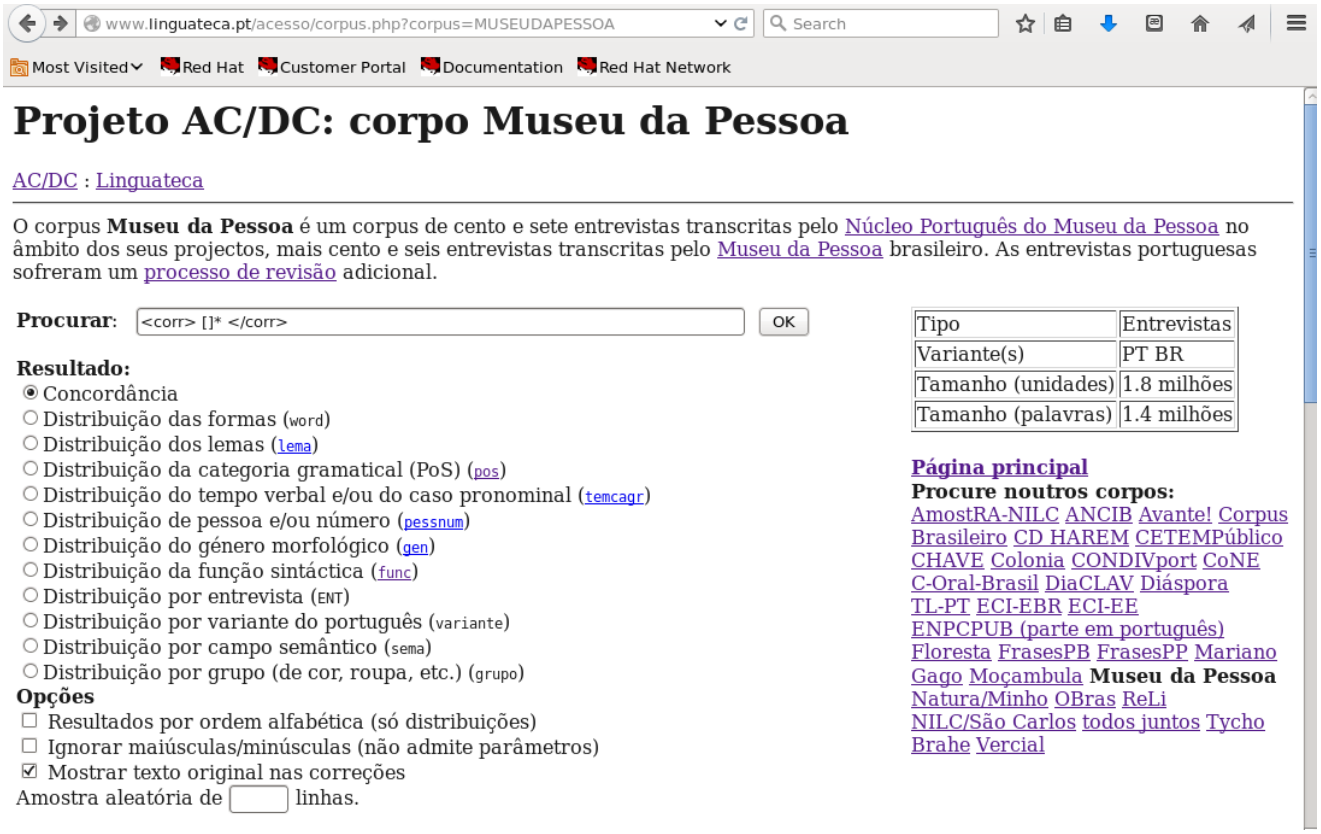
A nossa visão da língua portuguesa

- Parece-nos triste separar as variantes da língua como línguas diferentes, como por exemplo na *Gramática pedagógica do português brasileiro* de Bagno (2012), e não nos parece que a linguística tenha alguma coisa a ganhar com isso, e muito menos a lusofonia.
- Em relação ao processamento computacional da língua portuguesa, reputamos de muitíssimo útil termos criado recursos e programas que lidassem com as várias variantes, como fizemos durante mais de quinze anos na Linguateca.
- Isso não quer dizer que não achemos extremamente relevante estudar o português em todas as suas variedades. Não consideramos, obviamente, correto escamotear as diferenças que existem.

2. Questões metodológicas

- Formas diferentes de textos para diferentes usos. Sabemos que "oral transcrito" é uma família de entidades.
- O nosso objetivo era criar um texto para "publicação", no sentido de que repetições ou hesitações ou fragmentos ininteligíveis não prejudicassem a compreensão das histórias contadas, e que os termos mal grafados ou pronunciados não evitassem encontrar esses assuntos. (Visto que apenas um traço dialetal, *num* por não, tinha sido codificado pelos transcritores, resolvemos desfazê-lo.)
- A revisão não significou contudo que se perdesse a forma original de transcrição: mantivemo-la "escondida", mas acessível, no corpo (exceto as correções simplesmente ortográficas, ou seja, os erros introduzidos pelo transcritor)

Demonstração



3. Os resultados: primeira impressão

Os resultados da nossa revisão foram ao encontro do que dissemos antes sobre a língua portuguesa como um todo, embora tenhamos ficado surpreendidas pela variedade de problemas e de erros encontrados.
De facto, muitas das propriedades comumente elencadas pela variante brasileira para mostrar o seu afastamento da variante de Portugal foram encontradas em falantes idosos e pouco escolarizados do Norte do país... tradicionalmente mais conservador em termos linguísticos.

- a posição dos clíticos, a falta de mesóclise
- a falta de preposição nas relativas
- a falta de concordância ou o uso de uma concordância não padrão
- diferentes preposições
- *depois* com oração finita

Os resultados: primeira impressão

Outras questões de transcrição não têm a ver com a variante nem com a língua, mas com a dificuldade de interpretar nomes próprios (estrangeiros ou nacionais), quer por os falantes os terem travestido, quer por os transcritores os desconhecerem, ou ambas as razões:

- *Papa Piedoso* para *Papa Pio XII*,
- *Linhais da Serra* para *Unhais da Serra*
- *Passo Sousa Alves* para *professor Sousa Alves*
- ...

Outros ainda prendem-se com a interpretação correta do que foi dito.

A revisão pressupõe interpretação: um exemplo

Os operários devem saber o que podem exigir na empresa sem que a empresa vá abaixo, porque os patrões estão a explorar os operários quando paga a menos é porque não pode pagar.

1. Os operários devem saber o que podem exigir na empresa sem que a empresa vá abaixo, porque os patrões estão a explorar os operários quando pagam a menos e é porque não podem pagar.
2. Os operários devem saber o que podem exigir na empresa sem que a empresa vá abaixo, porque os patrões estão a explorar os operários... Quando (o patrão) paga a menos, é porque não pode pagar.

A revisão pressupõe interpretação: outro exemplo

Estes restaurantes já são um bocadinho antigos, embora estejam mais embelezados, não é, talvez consoante vai subindo vão ficando mais embelezados, mas digo-lhe uma coisa: já havia estes restaurantes, havia na mesma.

1. talvez consoante (os restaurantes) **vão** subindo (de categoria?) vão ficando mais embelezados...
2. consoante (o senhor) vai subindo vão ficando mais embelezados

Descrição quantitativa: tamanho do material

- Ao todo fizemos **1528** alterações em que mantivemos o original. (Repetimos que as simples correções ortográficas, como de *heide* para *hei-de* ou de *à* para *há*, não foram contabilizadas.)
- O texto (105 entrevistas) que foi alvo de correção continha 346 mil unidades (palavras e sinais de pontuação), correspondendo a 22.853 frases.
- Distribuição das correções por entrevista: Houve 89 entrevistas que sofreram pelo menos uma correção, sendo a E074 (249), a E020 (120) e a E041 (118) as que tiveram mais correções.

Descrição quantitativa: quanto às correções

As correções mais comuns foram:

- *prontos* para *pronto*
- *num* para *não*
- *era* para *eram*
- remover *que*
- *é* para *são*
- *tem* para *há*
- *diz* para *disse*
- *ao* para *no*
- *descrevia* para *descreveria* (na pergunta do entrevistador)
- remover *a*
- *lembra-me* para *lembro-me*
- remover *atrás*

Descrição quantitativa: quanto aos erros lexicais

Exemplos de "erros" lexicais ou morfológicos

- azilhargas, argelianos, imediático, asterose, escandinávios, trupedeados
- pequeninha, auguinha, perpetuzinhos, pouquexinhos
- prume, safes, parenta
- furecidade, desconfrar, antão, trás-dos-montes
- interviu, bebeides, comedes, deziam, camurflada
- mandavam-los, manifestaria-se, nos as, deveria-me acompanhar, trazia-nas, disse-lo
- consta-se, carrava, tinham quem lhes socorresse, abono-lhe
- a gente semos, deviam haver,
- mais grandes
- caneta permanente

4. Comentários finais

- Foi muito interessante observar a língua em uso sem passar pela monitorização da escrita. Claramente a questão da falta de concordância e da variabilidade da posição dos clíticos – muito bem documentados para o PB e para o PM – são casos que são comuns a todas as variantes do português.
- Não é fácil delimitar até onde se revê e que marcas do oral se devem manter, para manter a naturalidade e a fluidez da entrevista, mas parece óbvio que há diferentes níveis de autenticidade que se podem defender. O que é um texto (final)?
- Para tipos de estudos linguísticos diferentes, diferentes versões seriam preferíveis. Mantendo ambas, permitimos pelo menos dois tipos de estudos: o dos desvios à norma (escrita), e o do conteúdo e forma normalizados.

Trabalho futuro?

- Nova revisão para tratar das questões que se prendem com o discurso direto e indireto
- Anotação do tipo de correção para permitir uma quantificação mais fina: quantas correções de clíticos? Quantas de concordância? etc...
- Anotação no corpo da idade e género dos falantes, neste momento apenas acessível de `http://www.linguateca.pt/acesso/metadadosMP.html`

Referências mencionadas nesta apresentação

- Almeida, J. João, J. Gustavo Rocha, P. Rangel Henriques, Sónia Moreira & Alberto Simões. "Museu da Pessoa – arquitectura", *ABAD - Associação Bibliotecários e Arquivistas*, 2000.
- Bagno, Marcos. *Gramática pedagógica do português brasileiro*. Parábola Editorial, 2012.
- Bick, Eckhard. *The Parsing System Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Aarhus University Press, 2000.
- Santos, Diana. "Corpora at Linguateca: Vision and roads taken", in Tony Berber Sardinha & Telma de Lurdes São Bento Ferreira (eds.), *Working with Portuguese Corpora*, Bloomsbury, 2014, pp. 219-236.
- Simões, Alberto & Diana Santos. Nos bastidores da Gramateca: uma série de serviços". *Workshop on Tools and Resources for Automatically Processing Portuguese and Spanish, at PROPOR 2014, São Carlos, Brazil*, 2014, pp. 97-104.