Heloísa P. Vale
Maryualê M. Mittmann
Priscila O. Côrtes

VALIDAÇÃO DA TRANSCRIÇÃO E DA ANOTAÇÃO PROSÓDICA DE UM *CORPUS* ORAL

Objetivos

1. Apresentar o processo de validação da segmentação prosódica do C-ORAL-BRASIL.
   a) Treinamento dos transcritores.
   b) Validação prévia.
   c) Validação final.

Objetivos

2. Apresentar o processo de validação do conteúdo das transcrições do C-ORAL-BRASIL.
   a) Critérios de transcrição.
   b) Validação inicial.
   c) Validação final.

O C-ORAL-BRASIL

- *Corpus* de fala espontânea da variedade mineira (região metropolitana de BH), projeto coordenado por Tommaso Raso e Heliana Mello na UFMG.
- Macrosseção informal:
  - 139 textos (105 familiares e 34 públicos);
  - 208.130 palavras;
  - 21h 8min de gravação em situações naturais;
  - 1/3 monólogos, 1/3 diálogos e 1/3 conversações.
- Etapas de construção:
  - Gravação, transcrição, revisão, alinhamento (e revisão), revisão segmental, etiquetagem morfossintática, etiquetagem informacional; além das validações.

Revisão bibliográfica

- **Validação**: verificação de um *corpus* em relação a um conjunto de especificação técnicas preestabelecidas para sua construção (Schiel *et al*., 2004; van den Heuvel, 2000; van den Heuvel *et al.*, 2008).

- **Critérios**: especificações do *corpus* em conjunto com uma margem considerada aceitável para cada elemento especificado.

Revisão bibliográfica

Tipos de validação

- **Interna**: controle de qualidade realizado durante ou após a produção, conduzida pela própria instituição produtora do *corpus*.
- **Externa:** realizada por terceiros, sem vínculo com os desenvolvedores do *corpus*.

(Schiel *et al*., 2004)

Revisão bibliográfica

Tipos de validação

- **Formal:** automática, checa a codificação dos textos e simbologia utilizada.
- **De conteúdo:** manual, correção e acurácia em relação ao áudio, e se há desvios nas fronteiras da segmentação.

(Forsøe; Monachini, 2004; van den Heuvel *et al.*, 2008; Schiel *et al.*, 2004)

VALIDAÇÃO DA SEGMENTAÇÃO PROSÓDICA

Validação da segmentação: Metodologia

- Objetivo: uniformizar a anotação da segmentação prosódica no *corpus*.
- Formação dos transcritores.
  - Grupo 1 e Grupo 2.
  - Treinamento: testes e *feedback*.
- Validação prévia e validação final.
- Acordo entre anotadores: teste Kappa (Fleiss, 1971).

Validação da segmentação: Metodologia

- Simbologia da anotação da segmentação prosódica

Símbolo	Valor
//	Quebra prosódica com valor terminal. Delimita enunciados concluídos.
+	Quebra prosódica com valor terminal, mas indica que o eunciado não foi concluído.
/	Quebra prosódica com valor não terminal, indica fronteira de unidade prosódica dentro do enunciado.
[/nº]	*Retracting*, valor não terminal, "nº" indica o número de palavras envolvidas no fenômeno.

(Moneglia; Cresti, 1997; Raso; Mello, 2010)

Segmentação prosódica: Resultados

Validação prévia:

Tipo de acordo	Grupo 1*		Grupo 2**	
	Diálogo	Monólogo	Diálogo	Monólogo
Acordo geral	**0,78**	**0,76**	**0,77**	**0,82**
Quebras terminais	0,87	0,71	0,85	0,83
Quebras não terminais	0,58	0,66	0,66	0,75
Ausência de quebra	0,84	0,86	0,81	0,87

* Resultados após 3 testes.

** Resultados do diálogo após 5 testes e do monólogo após 8 testes.

Segmentação prosódica: Resultados

Validação final:

Tipo de acordo	Grupo 1		
	Geral	**Diálogo**	**Monólogo**
Acordo geral	**0,86**	**0,86**	**0,85**
Quebras terminais	0,87	0,87	0,86
Quebras não terminais	0,78	0,78	0,78
Ausência de quebra	0,91	0,91	0,90

Mais detalhes em Raso e Mittmann (2009).

VALIDAÇÃO DA TRANSCRIÇÃO

Critérios de transcrição

- Base ortográfica com implementação de alguns critérios não ortográficos.
- Critérios não ortográficos pretendem capturar fenômenos de gramaticalização ou lexicalização.
- Necessidade de equilíbrio entre captura dos fenômenos, legibilidade do texto e factibilidade da transcrição.

(Mello; Raso, 2009)

Exemplos

Aférese

*GER: aquela loucura toda / então / quer dizer / ser humano não **güenta** isso não / uai //

Cliticização

*MAI: o diâmetro **dea** deve dar uns [/1] uns quarenta cinqüenta centrímetro de [/1] de &s [/2] de grossura / o diâmetro **dela** //

Negação

*TER: lado des também **nũ** é rico **não** //

Preposições reduzidas

*JMA: toma dedeira **p'** cê dormir //

Validação da transcrição: Metodologia

- Objetivos: estabelecer o nível de confiabilidade das transcrições e melhorar a versão final do *corpus*.
- Tipos de erro validados:
  - Transcrição incorreta (grafia ou acurácia).
  - Inserção de palavras.
  - Deleção de palavras.
  - Aplicação inadequada dos critérios de transcrição.

Validação da transcrição: Resultados

Tipo de erro	**Validação inicial***		**Validação final****	
Todos os erros	**140/7484**	**1,87%**	**67/8243**	**0,81%**
Erros gerais	104/6319	1,65%	55/6124	0,90%
Transcrição incorreta	45/6319	0,71%	23/6124	0,38%
Inserção de palavra	19/6319	0,30%	19/6124	0,31%
Deleção de palavra	40/6319	0,63%	13/6124	0,21%
Aplicação inadequada de critérios de transcrição	37/1165	3,18%	12/2119	0,57%

* Realizada por dois transcritores.

** Realizada por um transcritor.

Considerações finais

Validação da segmentação

- O percentual de acordo obtido na validação final indica confiabilidade de 88% a 93% dos casos (intervalo de confiança de 95%).
- A anotação foi aplicada de modo uniforme, mesmo sendo realizada por transcritores diferentes.
- A validação prévia resultou em ganho de qualidade das transcrições.

Considerações finais

Validação da transcrição

- O resultado da validação final indica correção de 98,9% a 99,3% das palavras (intervalo de confiança de 95%).
- Alta precisão das transcrições do C-ORAL-BRASIL em relação à aplicação dos critérios ortográficos e não ortográficos.
- A metodologia de validação contribui de forma efetiva para o fortalecimento da Linguística de Corpus no Brasil.

Referências

FLEISS, J. L. (1971). Measuring nominal scale agreement among many raters. Psychological Bulletin, v. 76, p. 378-382.

FORSØE, H.; MONACHINI, M.; (2004). ELRA Validation Methodology and Standard Promotion for Linguistic Resources. *LREC 2004. Anais...* p. 941-944. Lisboa. Disponível em: <http://www.lrec-conf.org/proceedings/lrec2004/pdf/553.pdf>.

MELLO, H.; RASO, T. (2009). Para a transcriçao da fala espontânea: o caso do C-ORAL-BRASIL. *Revista Portuguesa de Humanidades*, v. 13, n. 1, p. 153-178. Disponível em: <http://www.ufjf.br/revistaveredas/files/2009/11/ARTIGO-Tommaso-Raso-e-Heliana-Mello.pdf>.

MONEGLIA, M.; CRESTI, E. (1997). L'intonazione e i criteri di trascrizione del parlato adulto e infantile. In: BORTOLINI, U.; PIZZUTO, E. *Il Progetto CHILDES Italia*. Pisa: Del Cerro, p. 57-90.

RASO, T.; MELLO, H. (2010). The C-ORAL-BRASIL corpus. In: MONEGLIA, M.; PANUNZI, A. (Eds.). *Bootstrapping Information from Corpora in a Cross-Linguistic Perspective*. Università degli studi di Firenze, Biblioteca Digitale.

RASO, T.; MITTMANN, M. M. (2009). Validação estatística dos critérios de segmentação da fala espontânea no corpus C-ORAL-BRASIL. *Revista de Estudos da Linguagem*, v. 17, n. 2, p. 73-91. Disponível em: <http://relin.letras.ufmg.br/revista/upload/17-2_04.pdf>.

SCHIEL, F. *et al.* (2004). The Validation of Speech Corpora. Disponível em: <http://www.phonetik.uni-muenchen.de/forschung/BITS/TP2/Cookbook/>.

van den HEUVEL, H. (2000). The Art of Validation. *The ELRA Newsletter*, v. 5, n. 4, p. 4-6. Paris. Disponível em: <http://www.elra.info/nl/newsletters/V5N4.pdf, 2000>.

van den HEUVEL, H. *et al*. (2008). Validation of spoken language resources: an overview of basic aspects. *Lang. Resources & Evaluation*, 42, p. 41-73.