Análise dos testes estatísticos Kappa dos segmentadores do corpus C-ORAL-BRASIL

Introdução

Este trabalho tem como objetivo analisar a segmentação do *corpus* C-ORAL-BRASIL (Raso e Mello, 2012) com base nos diversos testes Kappa realizados ao longo da formação dos segmentadores. A segmentação foi feita com base nos mesmos critérios do C-ORAL-ROM (Cresti e Moneglia, 2005), definidos na *Language into Act Theory* (Cresti, 2000; Raso, 2012; Moneglia e Raso, 2014). A segmentação é realizada por enunciados e unidades tonais. Define-se enunciado como a menor unidade pragmaticamente autônoma do fluxo da fala (Cresti, 2000; Cresti e Gramigni, 2004).

Com esse assunto teórico, foram realizadas, ao longo do tempo, diversas formações de segmentadores para os *corpora* da coleção C-ORAL, e agora também para um minicorpus extraído do Santa Barbara Corpus (Du Bois et. al., 2000-2005) ressegmentado e etiquetado informacionalmente no Laboratório de Estudos Empíricos e Experimentais da Linguagem (LEEL) da UFMG (Amorim, Ramos e Raso, em preparação). A segmentação, portanto prevê que o fluxo da fala seja segmentado em unidades tonais que podem ter valor conclusivo (e nesse caso são também fronteira de enunciado) ou não conclusivo (e nesse caso são em princípio fronteiras de unidade informacional e segmentam internamente o enunciado). Para realizar a segmentação, os segmentadores se baseiam na percepção de quebra prosódica, que indica fronteira de unidade tonal, e na percepção do valor conclusivo ou não conclusivo da quebra. A fronteira prosódica com valor terminal é marcada com uma barra dupla (//); a fronteira com valor não terminal com uma barra simples (/). Além desses, outros dois símbolos marcam respectivamente a fronteira de enunciado interrompido (+), que é, portanto uma fronteira terminal, e a fronteira de unidade tonal de retracting ([/]), que é portanto uma fronteira não terminal.

Por tratar-se de uma atividade dependente da percepção humana, é possível que, mesmo com um intenso treinamento, ocorram erros na segmentação prosódica do fluxo da fala, Por este motivo, a validação torna-se necessária para o caso de *corpora* orais. O teste estatístico Kappa (Cohen, 1960; Fleiss, 1971) é realizado, visando avaliar o coeficiente de concordância na delimitação de enunciados e de unidades tonais. O

resultado obtido varia entre 0 e 1, onde 0 indica desacordo total e 1 indica acordo total. Assim, quanto maior o for o valor de Kappa encontrado, maior será a concordância.

A versão final do C-ORAL-BRASIL foi validada com base em um teste Kappa, que como resultado obteve o valor de 0.86 (0.87 para as fronteiras terminais e 0.86 para as não terminais), ou seja, um valor julgado excelente. Para a análise da formação dos segmentadores e das modalidades de validação, veja-se Raso e Mittmann (2009), Moneglia et al. (2010), Mello et al. (2012). Ao longo do processo de formação foram tabulados os resultados de muitos testes Kappa, sejam aqueles que serviram para a validação final do *corpus*, sejam aqueles que decretaram concluída a formação dos segmentadores (sempre com Kappa superior a 0.8), e, ao mesmo tempo, os testes também serviram para identificar a natureza dos desacordos perceptuais.

Objetivos

Essa grande quantidade de material constitui um conjunto de dados extremamente rico para ser analisado. O objetivo desse trabalho é analisar os casos de desacordo nos diferentes testes e identificar os possíveis motivos que levaram ao desacordo na anotação. Ademais, visa-se propor uma possível classificação das causas dos desacordos, ou seja, dos fatores de natureza perceptual ou cognitiva que induzem mais facilmente os segmentadores a tomar decisões diferentes. De fato, esse trabalho se insere em um projeto que visa a identificar os parâmetros acústicos associados à percepção de quebra. Para isso, será elaborado um software capaz de extrair as configurações de parâmetros acústicos em volta das posições marcadas como fronteiras pelos segmentadores. Portanto, como o input do software será os trechos segmentados perceptualmente, é importante conhecer melhor os contextos de desacordos e suas possíveis causas, de modo a prever a probabilidade de erro humano nos dados fornecidos ao software e de modo a comparar as configurações acústicas que o software fornecerá com os pontos mais críticos do ponto de vista perceptual.

Metodologia

Para o cálculo do teste estatístico Kappa, os transcritores realizaram a anotação das quebras prosódicas perceptíveis nos textos a eles oferecidos, que já estavam transcritos, mas subtraídos da anotação da segmentação. Assim, durante a realização do teste, os transcritores precisavam apenas diferenciar as quebras terminais

(concluídas e interrompidas) das quebras não terminais (não terminais e *retractings*), no prazo de três dias sem discutir a segmentação com outras pessoas.

Por fim, o cálculo do Kappa foi realizado através do ambiente para computação estatística R (R Development Core Team, 2010). Ao término teste, as quebras terminais e não terminais que geraram desacordos, quanto à anotação da segmentação, foram ouvidas e analisadas diversas vezes. As divergências foram estudadas caso a caso, visando entender os motivos pelos quais os transcritores segmentaram de maneiras distintas. Posteriormente, foram estabelecidas quais foram as possíveis razões perceptuais ou cognitivas que causaram as diferenças na segmentação.

Resultados

Os resultados indicam que a marcação incorreta na segmentação, em relação às quebras terminais e não terminais, decorre de fatores tanto de natureza cognitiva quanto de natureza perceptual. Os erros podem ser categorizados com base em causas possíveis, entre as quais as principais e mais frequentes parecem ser: (i) quebra de natureza sintática, mas não marcada pela prosódia. Nesses casos, o segmentador abandona o critério perceptual e segue um critério de outra natureza; (ii) fronteira pragmática. Isso acontece principalmente em começo de enunciado depois de um item lexical ao qual pode ser atribuído valor de marcador discursivo. Muitos dos casos de desacordo são relativos a esse tipo de posição; (iii) ênfase, ou seja, proeminências prosódicas de natureza local, que não têm como escopo a unidade inteira. Nesses casos, parece que à percepção de uma perturbação prosódica ao longo do fluxo de fala é às vezes atribuído o valor de quebra; (iv) coarticulação, hipoarticulação e aumento da velocidade de fala em posição pré-fronteiriça. Nesses casos aparecem muitos desacordos com relação ao valor terminal ou não terminal da fronteira; (v) fatores de natureza rítmica. Às vezes, há sequências de fronteiras (normalmente não terminais) distribuídas em poucas palavras. Ou seja, temos séries de unidades pequenas. Nesses casos, ainda mais se associados à taxa de articulação alta, podemos ter desacordos no número das quebras, principalmente na primeira fronteira.

Referências

AMORIM, F.; RAMOS, A.; RASO T. A resegmented and informationally tagged minicorpus extracted from the SBC (em preparação)

COHEN, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, v. 20, p. 37-46.

CRESTI, E. (2000). *Corpus di italiano parlato.* Firenze: Accademia della Crusca, v.1.

CRESTI, E. & GRAMIGNI, P. (2004) Per una linguistica corpus based dell‟italiano parlato: le unità di riferimento. In: F. Albano Leoni , F. Cutugno, F. Pettorino, M. Savy & R. Savy (eds.), *Atti del Convegno Nazionalle "Il Parlato Italiano",* CD-ROM (pp. 1 – 26). Napoli: M. D'Auria.

CRESTI, E.; MONEGLIA, M. (Eds.) (2005). *C-ORAL-ROM*: Integrated Reference Corpora For Spoken Romance Languages. Amsterdam: John Benjamins, 304 p. (Inclui Multimedia Corpus em DVD).

DUBOIS, J. W. et al. (2000). *Santa Barbara corpus of spoken American English* - Part 1. Linguistic Data Consortium.

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

MELLO, H. et al (2012). Transcrição e segmentação prosódica do corpus C-ORAL-BRASIL: critérios de implementação e validação. In: RASO, T. & MELLO, H. (orgs.). *C-ORALBRASIL I: Corpus de referência do português brasileiro falado informal.* Belo Horizonte: Editora UFMG.

MONEGLIA M.; RASO T.; MITTMANN, M. M.; RIBEIRO MELLO, H. (2010) "Challenging the Perceptual Relevance of Prosodic Breaks in Multilingual Spontaneous Speech Corpora: C-ORAL-BRASIL / C-ORAL-ROM." In: *Prosodic Prominence Perceptual and Automatic Identification* - Speech Prosody 2010 Satellite Workshop. Chicago: Université de Neuchâtel

MONEGLIA, M. AND RASO T. (2014) "Appendix". In *Spoken Corpora and Linguistic Studies*, Raso, Tommaso and Heliana Mello (eds.), 468–495

R DEVELOPMENT CORE TEAM. (2010). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Disponível em: <http://www.r-project.org>.

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>.

RASO, T.; MELLO, H. (2012). *C-ORAL-BRASIL I: corpus de referência do português brasileiro falado informal.* Belo Horizonte: Editora UFMG, 332 p.

RASO, T.; (2012). O corpus C-ORAL-BRASIL. In: RASO T.; MELLO H. (Org.). In: *C-ORAL-BRASIL I. Corpus de referência do português brasileiro falado informal.* Belo Horizonte: Editora UFMG, p. 55-90.