DOI: 10.17851/2237-2083.23.3.883-922

O papel da pausa na segmentação prosódica de *corpora* de fala

The role of pause in the prosodic segmentation of spoken corpora

Tommaso Raso
Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, Minas Gerais, Brasil
tommaso.raso@gmail.com

Maryualê Malvessi Mittmann
Universidade do Sul de Santa Catarina (UNISUL), Tubarão, Santa Catarina, Brasil
maryuale@gmail.com

Anna Carolina Oliveira Mendes
Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, Minas Gerais, Brasil
annacarol.om@gmail.com

Resumo: O trabalho apresenta uma análise da pausa silenciosa enquanto critério de segmentação da fala em unidades comunicativamente autônomas. O objetivo deste trabalho é contextualizado com uma discussão sobre a unidade de referência da fala e a noção de pausa. Argumenta-se a favor de uma segmentação da fala em unidades pragmáticas de natureza acional, delimitadas no fluxo da fala por meio de fronteiras prosódicas. Realizou-se análise estatística em amostra aleatória de fala espontânea extraída do *corpus* C-ORAL-BRASIL. Por intermédio de um modelo de regressão não linear, procurou-se identificar durações de pausas consistentemente preditivas de fronteiras de unidades de referência. Os resultados mostram que não há nenhum valor de duração de pausa que possa ser utilizado como critério confiável para a segmentação da fala em unidades comunicativamente autônomas.

Palavras chave: fala; corpora; segmentação; unidade prosódica; ilocução.

Abstract: This work presents an analysis of silent pauses when applied as criterion for the segmentation of speech into communicatively autonomous units. The aim of this work is contextualized in a discussion about the unit of reference for speech and the notion of pause. We argue that speech should be

parsed into pragmatic units of actional nature, which are signaled within the speech flow by prosodic boundaries. We performed a statistic analysis on a random spontaneous speech sample extracted from C-ORAL-BRASIL corpus. Non-linear regression models were applied so as to identify pause durations that could consistently predict unit boundaries. Results show that there isn't any pause duration value that could be used as a reliable criterion for the segmentation of speech into communicatively autonomous units.

Keywords: speech; corpora; segmentation; prosodic unit; illocution.

Recebido em: 25 de setembro de 2015.
Aprovado em: 23 de outubro de 2015.

1 Introdução

Investigamos a relação entre pausas por um lado e fronteiras de enunciados e unidades tonais por outro, conforme definidos adiante. Essa investigação baseia-se no *corpus* C-ORAL-BRASIL (RASO; MELLO, 2012) e insere-se em uma discussão acerca da unidade de referência da fala, ressaltando a importância de uma segmentação da fala adequada para representar fenômenos próprios dessa diamesia[1] e questionando diferentes propostas desenvolvidas na literatura.

Primeiramente, discutimos brevemente o conceito de unidade de referência da fala e as principais propostas. Em seguida, aprofundamos em uma das propostas, também apresentando a perspectiva da *Language into Act Theory*, que norteia a elaboração dos *corpora* de fala espontânea C-ORAL-ROM e C-ORAL-BRASIL. Depois, apresentamos a arquitetura do C-ORAL-BRASIL; e, por fim, apresentamos a metodologia e os resultados da pesquisa.

[1] Para o conceito de *diamesia*, veja-se Berruto (1993) e Rossi (2011).

2 O problema da segmentação da fala

Algumas características tornam a fala intrinsecamente distinta da escrita e, portanto, a nosso ver, não é possível transpor categorias analíticas da escrita para a fala. A discussão sobre diferenças entre fala e escrita é ampla. Remetemos a Raso (2013) para uma síntese e a proposta que embasa este trabalho. Apenas lembramos que essas diferenças ocorrem porque a fala e a escrita são transmitidas por meios e canais diferentes, e são decodificadas por sentidos diferentes. Isso gera consequências pragmáticas e linguísticas muito distintas na organização comunicativa. Característica própria da fala, inevitavelmente ausente na escrita, é a transmissão da informação pelo canal sonoro. Desse canal faz parte não somente o conteúdo segmental (articulado, percebido e decodificado de modo muito diferente da realização e decodificação dos grafemas), mas também a prosódia, com sua enorme bagagem de informações que, na escrita, desaparece ou é veiculada por procedimentos físicos e cognitivos diferentes. Isso torna impossível a transposição da fala para a escrita sem que haja, de fato, uma perda enorme de informações e a transferência de algumas delas para categorias diferentes. Por isso, consideramos impossível estudar adequadamente a fala recorrendo exclusivamente a transcrições, sem considerar as informações veiculadas por meio do som, principalmente, os efeitos da prosódia.

A necessidade de se recorrer constantemente ao áudio para avaliar os fenômenos linguísticos da fala impõe que os *corpora* disponibilizem o alinhamento do texto ao som. O alinhamento evidencia ainda mais a necessidade de definir o domínio mínimo dos principais fenômenos linguísticos, ou seja, a unidade de referência da fala maior do que a palavra.

Naturalmente, é oportuno identificar tal unidade com base nos princípios que governam a comunicação oral, e não transpondo mecanicamente estruturas que funcionam na análise da escrita. A unidade de referência deve ser entendida como o âmbito de funcionamento das principais relações entre os elementos linguísticos, nas suas várias ordens: sintática, semântica ou pragmática.

Um exemplo pode esclarecer essa questão. Imaginemos a sequência *João vai pro Rio até amanhã*. Ela pode receber diferentes segmentações, entre as quais:

(i) *João vai pro Rio até amanhã //* (uma única unidade de referência);
(ii) *João vai pro Rio // até amanhã //* (duas unidades de referência);
(iii) *João // vai pro Rio até amanhã //* (também duas unidades de referência);
(iv) *João // vai pro Rio // até amanhã //* (três unidades de referência).

Junto à segmentação, atribui-se um valor acional às unidades. Em (i), podemos ter uma asserção, uma pergunta, uma manifestação de surpresa ou outras ações; em (ii), temos duas ações, a primeira poderia ser uma asserção, uma resposta, um pedido de confirmação ou outra, e a segunda, uma despedida, uma outra asserção, um outro pedido de confirmação ou outra; em (iii), também temos duas ações, mas relativas a conteúdos locutivos diferentes, podendo a primeira ser um chamamento, ou um pedido de confirmação, ou uma resposta ou outra, e a segunda, uma ordem, uma pergunta ou outra; em (iv), temos três ações, a primeira podendo ser um chamamento, um pedido de confirmação, uma expressão de surpresa ou outra; a segunda, podendo ser uma ordem, uma expressão de incredulidade, ou de surpresa ou outra; e a terceira, podendo ser uma despedida, uma asserção, uma pergunta ou outra.

Somente após identificarmos as unidades e atribuirmos a elas um valor acional podemos dizer se *João* é sujeito ou o que tradicionalmente é chamado de "vocativo", e se a forma verbal *vai* deve ser analisada como terceira pessoa do presente do indicativo ou segunda pessoa do imperativo. Parece-nos evidente que tanto a segmentação quanto a atribuição de valor acional são guiados pela percepção de marcas de natureza prosódica.

Frequentemente dá-se o nome de *enunciado* a essa unidade de referência, mas sob essa denominação são apresentados objetos diferentes. Na tradição pragmática (AUSTIN, 1962), o termo *enunciado* faz referência ao que é considerada a mínima unidade comunicativa,

capaz, portanto, de veicular uma mensagem mínima e autônoma, com um núcleo comunicativo de natureza acional. Como identificar e descrever as estruturas linguísticas que realizam essa função é ainda objeto de discussão. Quatro são as principais perspectivas.

2.1 Diversas visões sobre a unidade de referência da fala

2.1.1 A sentença falada

Na perspectiva sintaticista, define-se enunciado como uma "sentença falada". Consideramos as duas principais definições de sentença: (a) como predicação verbal relativa a um SN sujeito, hierarquicamente subordinado ao SV (HARRIS, 1962); (b) como a "máxima projeção do núcleo verbal" (CHOMSKY, 1970).

Com base no *corpus* LABLITA de italiano falado (CRESTI, 2000), Cresti e Gramigni (2004) observam que estruturas do tipo (a) constituem menos de 5% dos enunciados, ou seja, uma frequência muito baixa para ser considerada fenômeno relevante para a fala. Mello, Raso, Bossaglia e Santana (em preparação), analisando o *minicorpus* extraído do *corpus* C-ORAL-BRASIL (RASO, 2012a; MELLO, 2014; PANUNZI; MITTMANN, 2014), observam que em Português Brasileiro (PB) as predicações constituem cerca de 15% dos enunciados, mas aquelas preenchidas com um SN pleno e um SV pleno representam pouco mais de 1%. Cresti e Gramigni (2004) já haviam observado que no italiano a maioria das predicações são constituídas por sujeitos pronominais e / ou verbos de ligação. As predicações em PB são bem mais frequentes que em italiano, mas ainda aparecem em quantidade muito reduzida. Essas diferenças não devem ser atribuídas a características dos *corpora*, que são comparáveis, mas às características acentuais das duas línguas. O PB, principalmente na variedade mineira (representada no C-ORAL-BRASIL), é mais acentual do que o italiano, o que permite hospedar nas unidades tonais maior quantidade de material fonológico, possibilitando estruturas sintáticas mais complexas.

Estruturas do tipo (b) são consideravelmente mais comuns nos dois *corpora*, variando entre 62% e 70%, dependendo da língua e de fatores sociolinguísticos. Em consonância com o evidenciado por Biber *et al*. (1999) para o inglês, Cresti e Gramigni (2004) observam, no

italiano, que 38% dos enunciados são não verbais. Para o PB, Raso e Mittmann (2012) notam que 27% dos enunciados não apresentam qualquer forma verbal e apenas 55% apresentam forma funcionalmente verbal[2] como núcleo do enunciado. A definição de enunciado como máxima projeção de V mostra-se, portanto, incapaz de explicar mais de um terço das unidades presentes em *corpora*.

Ambas as definições do enunciado como "sentença falada" deixam de fora um número grande demais de casos para serem consideradas satisfatórias na análise da fala.

2.1.2 O turno

Na perspectiva dialógica, adota-se o turno como unidade de referência. Em *corpora* de fala espontânea como o C-ORAL-ROM (CRESTI; MONEGLIA, 2005) para espanhol, francês, italiano e PE, o C-ORAL-BRASIL para o PB e o Santa Barbara *Corpus* (DU BOIS *et al.*, 2005) para o inglês americano, observa-se que essa delimitação corresponde a objetos muito heterogêneos. Turnos podem abranger uma única palavra (ou mesmo interjeição), ou estruturas extremamente complexas, com duração de muitos minutos. Devido a tamanha heterogeneidade, o turno, apesar de constituir uma unidade natural da fala, não é adequado para a delimitação da unidade mínima significativa maior do que a palavra.

Os exemplos a seguir ilustram sequências de turnos em uma conversação e um diálogo, e um fragmento de texto prevalentemente monológico.[3]

---

[2] Uma parte significativa dos enunciados com verbo pertencem a duas tipologias: a) enunciados em que o verbo aparece como expansão de um núcleo de outra natureza, por exemplo um SN modificado por uma relativa; b) formas verbais que retomam o verbo do enunciado anterior para expressar confirmação ou negação, como no exemplo fictício a seguir:
*AAA: você ligou para João //
*BBB: liguei //

[3] A sigla de três letras precedidas de asterisco e seguida de dois pontos indica o falante do turno; os números entre colchetes indicam o número dos enunciados no texto; os símbolos "<>" indicam sobreposição de fala; as barras duplas indicam final de enunciado; as barras simples indicam final de unidade tonal; o símbolo "+" indica enunciado interrompido; a barra simples entre colchetes e seguida por um

Ex. 1 - bfamcv05 [75-95][4]

áudio 1

*JOS: [75] <e as> mordomia que es têm //
*CAR: [77] na hora que fizer cinco / nós vamo parar cinco minutos / viu // [78] porque / <tem jeito não> //
*JOS: [79] <ou mais / né> // [80] ué / onde é que essa bola foi //
*CAR: [81] <machucou / ô> //
*JOS: [82] <no &go [/2] lá no quintal do vizinho> //
*CEL: [83] espim / cara //
*CAR: [84] cuidado aí que <tem> [/1] tem coisa aí //
*MAR: [85] <possível> / aí //
*CAR: [86] vai dar pra jogar //
*CEL: [87] não / vai / sô //
*JOS: [88] furou nada não //
*MAR: [89] caco de vidro //
*CEL: [90] não / foi <espinho mesmo> //
*CAR: [91] <nũ é não> / ali tem / ora-pro-nóbi //
*MAR: [92] <não / é de mão / meu> filho //
*CEL: [93] <tá a bosta / mesmo aqui o'> //
*JOS: [94] não // [95] tanto faz //

O exemplo 1 mostra uma alternância de turnos em uma conversação fortemente interativa, com turnos extremamente curtos: trata-se de quatro amigos jogando futebol. Em 89 palavras, temos 17

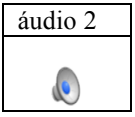
---

número indica *retracting* (o número indica a quantidade de palavras retratadas, ou seja virtualmente deletadas, pelo falante); o símbolo "&" indica palavra não terminada; o símbolo "hhh" indica riso ou tosse ou som sem valor linguístico; a sequência "&he" indica tomada de tempo, frequentemente chamada também de pausa preenchida.

[4] Cada exemplo é acompanhado da referência do texto e dos enunciados aos quais se refere, sempre relativos ao *corpus* C-ORAL-BRASIL. A sigla de identificação do texto é composta por uma letra que identifica a língua (em nosso caso sempre "b"), por três letras que identificam o contexto, se familiar-privado ("fam") ou público ("pub"), por duas letras que identificam a tipologia interacional, monólogo ("mn"), diálogo ("dl") ou conversação ("cv"), esse último entendido como um diálogo com mais de dois participantes, e por um número que identifica o texto dentro da categoria definida pela sigla. O(s) número(s) entre colchetes indicam o enunciado ou a sequência de enunciados.

turnos. Observe-se que apenas três foram segmentados em dois enunciados (delimitados pela barra dupla), enquanto 14 são formados por um único enunciado. Em exemplos assim, a coincidência entre turno e enunciado é quase total. Observe-se também que seis enunciados e cinco turnos não apresentam elemento verbal, e outros dois apresentam verbos não nucleares.

Ex. 2 - bpubdl03 [118-135]



*GUI: [118] volta aqui //
*GUI: [119] faz força //
*GUI: [120] mais //
*GUI: [121] beleza //
*GUI: [122] contrai o abdômen //
*GUI: [123] joga o tronco só um pouquinho pra frente //
*GUI: [124] aí //
*GUI: [125] beleza //
*GUI: [126] descansou //
*GUI: [127] vou baixar um pouquinho mais //
*GUI: [128] vai //
*GUI: [129] pera aí / deixa eu passar a faixa //
*GUI: [130] aí //
*GUI: [131] vai / força //
*GUI: [132] aqui //
*GUI: [133] pra frente // [134] isso //
*GUI: [135] pesado //

O exemplo 2 mostra uma alternância de 17 turnos do mesmo falante em 41 palavras. Trata-se de um diálogo entre *personal trainer* e cliente, durante uma sessão de ginástica. Aqui, não há alternância de falante; GUI produz turnos alternados a comportamentos não verbais do ouvinte. Evidentemente não podemos considerar esse um trecho monológico. O falante não segue um plano textual independente da interação. Ao contrário, o trecho é fortemente interativo, mas um dos participantes interage de maneira não verbal. Os enunciados do falante não podem, portanto, ser considerados como pertencentes ao mesmo turno, mas como turnos que se alternam a turnos não verbais, mas "agidos", pelo interlocutor. Observe-se que apenas um turno foi segmentado em dois enunciados e que oito turnos e nove enunciados não possuem verbo.

Ex. 3 - bfammn06 [6-17]

áudio 3


*JOR: [6] bom / eu tive / a minha / formação / profissional / dentro da / área de engenharia / depois que eu fiz escola técnica no Rio de Janeiro / e passei pela administração na Fundação Getúlio Vargas // [7] na + [8] já tem algum tempo que eu tô formado / naquela época / o mercado de trabalho era totalmente diferente de hoje // [9] hhh a [/1] as multinacionais estavam entrando dentro do país / e procuravam / alunos dentro das próprias universidades // [10] e assim eu iniciei minha vida profissional / na área técnica de engenharia elétrica // [11] um belo dia durante o almoço / o gerente de recursos humanos de uma multinacional / me informou que havia uma vaga na área comercial da empresa / e / se eu tinha interesse // [12] eu &fo + [13] eu &es + [14] informei a ele que eu tava preste a me formar / e / estava trabalhando dentro duma / área que eu gostava // [15] mas / ele me informou / que o salário seria quase o dobro do que eu ganhava // [16] e aquilo mexeu muito comigo // [17] e aí / eu consegui / a [/1] com a experiência que eu tinha dentro da multinacional / concorrer à vaga e &f [/1] isso me facilitou / e eu passei pra área comercial da empresa pra vender / disjuntores / transformadores / motores de / corrente contínua / corrente alternada / isoladores / e / relés de proteção secundária / e assim foi iniciando a minha vida comercial //

O exemplo 3 mostra apenas um turno de uma situação tendencialmente monológica (na fala espontânea informal não existe monólogo perfeito). Em 104 palavras, temos apenas um turno, segmentado em 17 enunciados. Observe-se que, com exceção de dois interrompidos, todos os enunciados possuem núcleo verbal.[5] A natureza do texto, a sua diafasia, se revela, portanto, decisiva na composição sintática dos enunciados.

Esses exemplos mostram como o turno, mesmo sendo uma unidade natural da fala, não serve para identificar a unidade de referência da fala, entendida como o âmbito no qual identificar uma intenção comunicativa mínima e autônoma do falante e as relações estreitas dos elementos linguísticos. Isso nos induz a procurar algo entre a palavra e o turno que constitua o domínio das principais relações linguísticas em uma unidade comunicativa mínima.

[5] Para uma análise comparativa da estrutura de textos dialógicos e monológicos, veja-se Raso e Mittmann (2012) e Mittmann (2013).

2.1.3 A pausa

Frequente é a definição de enunciado como sequência entre duas pausas do mesmo falante. Primeiramente, precisa-se fazer algumas considerações sobre o que entendemos por pausa. Comumente, distingue-se entre "pausa silenciosa" e "pausa preenchida" (ZELLNER, 1994; SORIANELLO, 2006; MERLO; BARBOSA, 2012). A pausa silenciosa seria constituída por um silêncio no fluxo locutivo; voltaremos a esse tipo de pausa, objetivo principal do trabalho. A pausa preenchida é considerada uma interrupção cognitiva que, entretanto, é realizada com alguma vocalização. Quem agrupa em um mesmo conceito os dois tipos de pausa, de fato, considera a vocalização um elemento superficial, que não merece um estatuto diferenciado por si só. A nosso ver, pausa silenciosa e pausa preenchida não podem ser tratadas como duas formas de uma única categoria. Tomar uma decisão definitória com base em uma suposta identidade cognitiva em presença de correlatos físicos diferentes não nos parece metodologicamente confiável.

A pausa silenciosa é uma interrupção do fluxo da fala, mesmo que provisória, identificável perceptualmente sem ambiguidade. Pode gerar efeitos involuntários (como perda do turno) ou efeitos voluntários de ordem comunicativa (como ressaltar um trecho de fala ao criar expectativa). A pausa preenchida é um fenômeno principalmente de disfluência, ainda mais se nessa categoria se inserem as repetições. Pausa silenciosa e pausa preenchida podem ou não se sobrepor funcionalmente. Mesmo certas expressões linguísticas podem ter funções parecidas com algumas das funções da pausa, silenciosa ou preenchida.

Consideramos importante distinguir entre o nível da identificação de fenômenos próprios da fala em termos de características físicas, e o nível da interpretação funcional desses fenômenos, que pode variar em contextos distintos. Assim, diferenciamos esses casos, deixando o termo "pausa" unicamente para aquela silenciosa. As vocalizações serão consideradas tomadas de tempo, marcadas em nossa transcrição com o símbolo "&he", e na etiquetagem funcional das unidades tonais com =TMT= (*time taking*); as repetições são consideradas casos de *retracting*; os *retractings* são correções que os falantes fazem, com ou sem repetições; em nossa notação, são marcadas com uma barra simples entre colchetes e um número que indica a quantidade de palavras retratadas; os

retractings são etiquetados como =EMP= (*empty*), para marcar que se trata de unidades informacionalmente vazias. Tanto as tomadas de tempo quanto as palavras retratadas devem poder ser isoladas na computação do número total das palavras, já que não possuem o mesmo valor informativo do resto (MONEGLIA; RASO, 2014; PANUNZI; MITTMANN, 2014).

Em vários arcabouços, as pausas são consideradas marca de fronteira de enunciado. De fato, uma pausa silenciosa igual a ou maior que 200ms foi considerada como fronteira em alguns *corpora* de fala, como no *Dutch Corpus* (BUHMANN *et al.*, 2002) e uma pausa igual a ou maior que 100 ms foi considerada fronteira em *corpora* japoneses (MARUYAMA, 2012, 2015). A vantagem disso é evidente: estabelecendo uma duração de silêncio como marca de fronteira, pode-se realizar a segmentação da fala automaticamente. Contudo, também as desvantagens são evidentes. Não há uma medida *a priori* ancorada a algum aspecto de natureza perceptual ou linguística que defina a duração mínima e máxima de um silêncio que possa, por si só, ser considerado fronteira de enunciado.

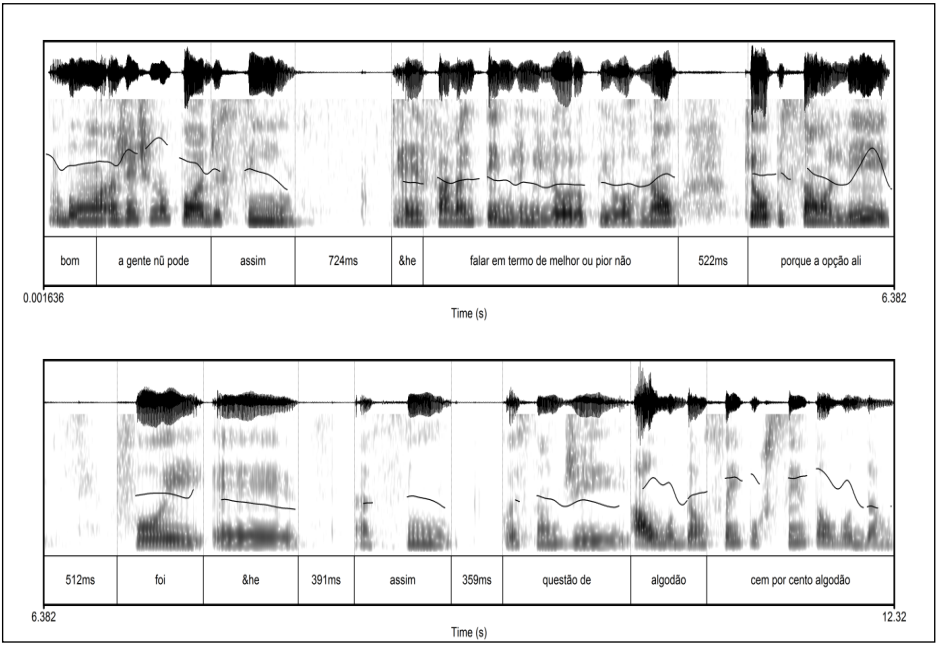
Os exemplos seguintes mostram casos de pausas, tomadas de tempo, *retractings* e coocorrência dos três. Neles as pausas são sempre maiores que 200ms e, à oitiva, resultam claramente não fronteiriças, no sentido de não separar duas unidades percebíveis como comunicativamente autônomas. Nesses casos, as pausas constituem apenas silêncios internos à mesma unidade comunicativa mínima. Outros exemplos mostram como unidades comunicativamente autônomas podem ocorrer na fala sem serem separadas por pausas. O intuito é duplo: mostrar que pausas silenciosas e preenchidas não podem ser consideradas duas formas do mesmo fenômeno e mostrar que a pausa não é nem suficiente nem necessária para marcar fronteira de enunciado.

Exemplo 4 - bfamcv19 [37]

áudio 4


*MAE: bom / a gente nũ pode / assim / &he / falar em termo de melhor ou pior não / porque a opção ali / foi / &he / assim / questão de / algodão / cem por cento algodão //

Figura 1 – Sinal de áudio, espectrograma, F0 e transcrição do exemplo 4



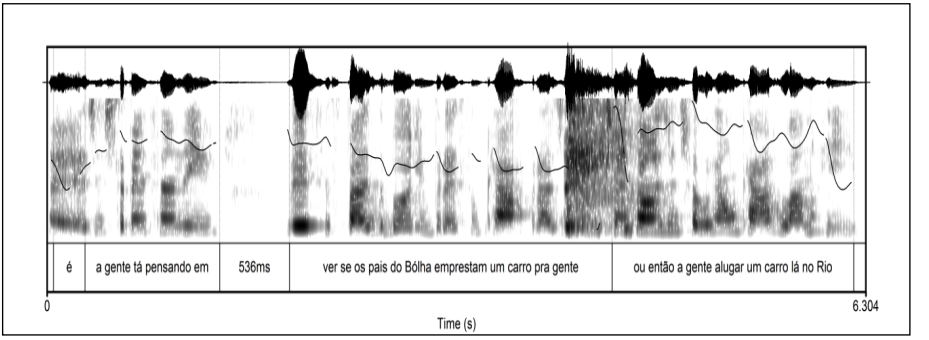
Esse enunciado apresenta pausas longas e tomadas de tempo: primeiro uma pausa de 724ms, seguida de uma rápida tomada de tempo; depois, uma segunda pausa de 522ms; uma terceira de 512ms, por sua vez, seguida de uma palavra (foi), uma tomada de tempo e uma pausa de 391ms, ainda seguida de uma palavra (assim) e uma outra pausa de 359ms. A rigor, ainda se segue uma pausa menor (147ms), após duas palavras (questão de) antes do fim do enunciado.

Exemplo 5 - bfamcv29 [37]

áudio 5

*ELI:é / a gente tá pensando **em / ver** se os pais do Bólha emprestam um carro pra gente / ou então a gente alugar um carro lá no Rio //

Figura 2 – Sinal de áudio, espectrograma, F0 e transcrição do exemplo 5



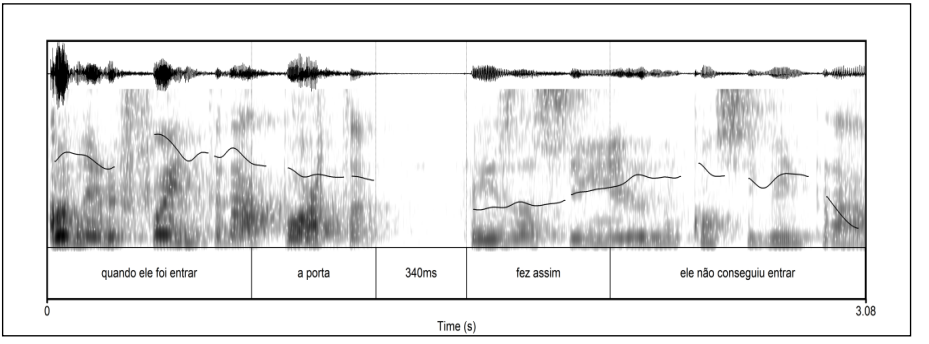
No exemplo 5, temos, entre outras, uma pausa de 536ms entre a preposição e o verbo por ela regido, evidenciando ainda mais que duração de pausa não é correlato confiável para determinar a existência de fronteira de unidade mínima comunicativamente completa, por estar em uma posição impossível de ser considerada fronteiriça do ponto de vista sintático, e não somente pragmático e prosódico.

Exemplo 6 - bfamdl14 [02]

áudio 6	áudio 6a

Quando ele foi entrar / **a porta / fez assim** / ele não conseguiu entrar //

Figura 3 – Sinal de áudio, espectrograma, F0 e transcrição do exemplo 6



No exemplo 6, uma pausa de 340ms está entre os elementos que,

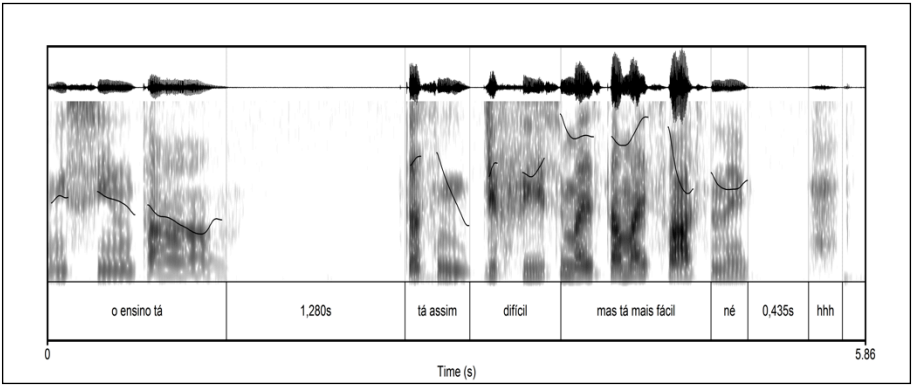
em uma análise sintática, seriam considerados o sujeito e o verbo. Mesmo em uma análise prosódica, essa posição não poderia ser considerada fronteira de enunciado, como mostram as oitivas do enunciado (áudio 6) e do trecho até a pausa (áudio 6a).

Exemplo 7 bpubdl11 [113]

áudio 7	áudio 7a	áudio 7b

o ensino tá [/1] tá assim / difícil / mas tá mais fácil / né hhh //

Figura 4 – Sinal de áudio, espectrograma, F0 e transcrição do exemplo 7



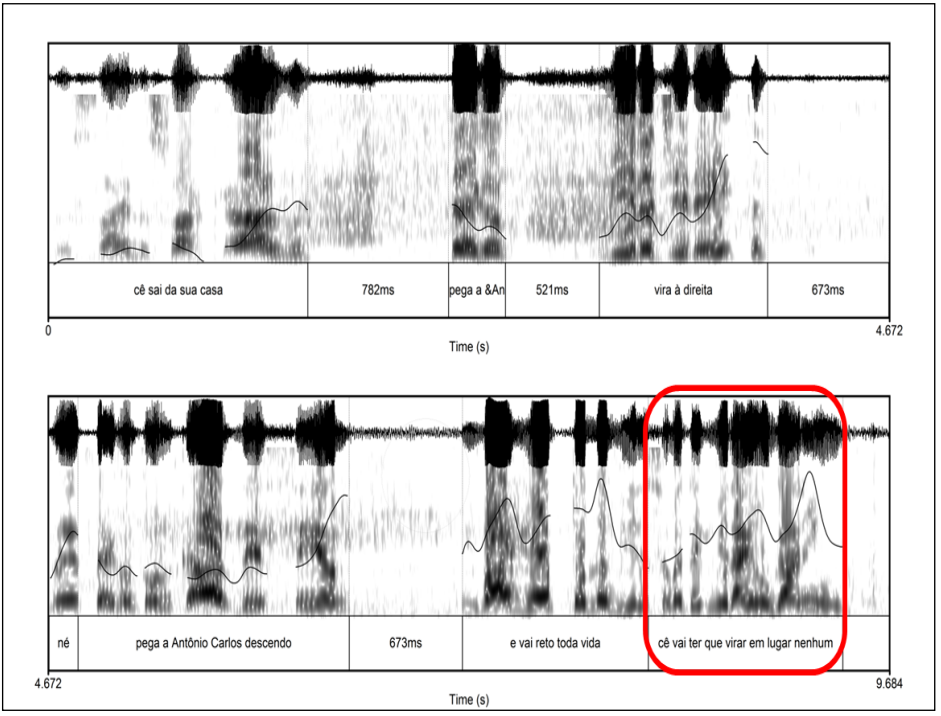
O exemplo 7, além da pausa, apresenta repetição de uma palavra. Aqui temos dois fenômenos não reduzíveis a um: a repetição, que poderia ter acontecido sem pausa (como mostra o áudio 7a editado), e o silêncio de 1,28s. Nesse caso também, a pausa não pode constituir marca de fronteira de enunciado, como mostra o áudio 7b.

Exemplo 8- bfamdl08 [34-35]

áudio 8

*AND: cê sai da sua casa / pega a &An [/3] vira à direita / é / pega a Antônio Carlos descendo / e vai reto toda vida // cê vai ter que virar em lugar nenhum //

Figura 5 – Sinal de áudio, espectrograma, F0 e transcrição do exemplo 8



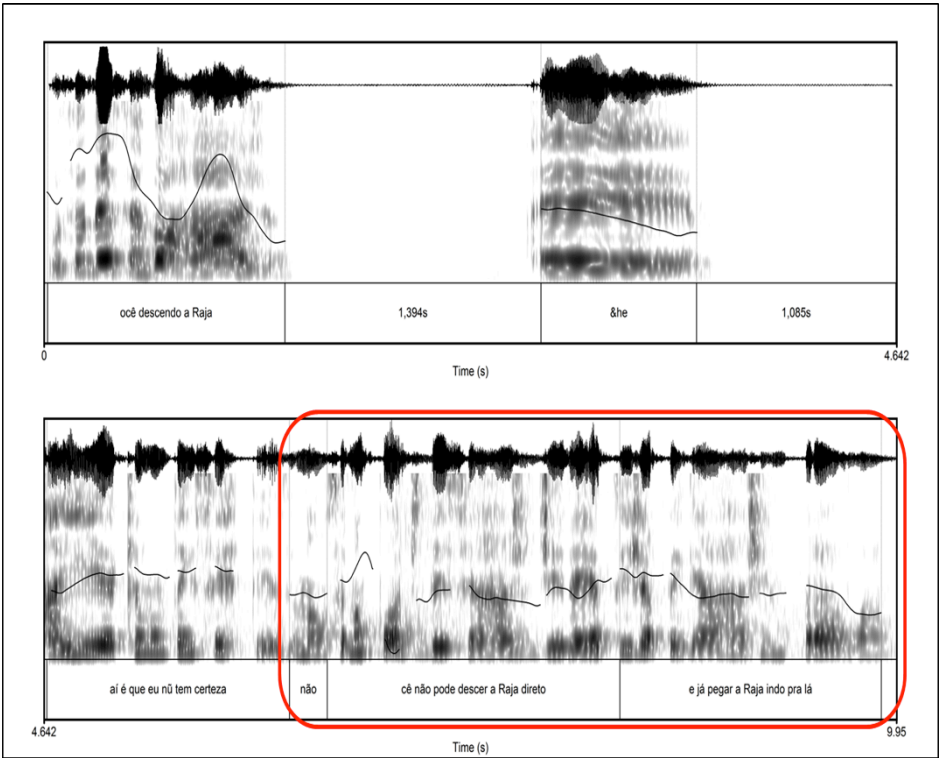
No exemplo 8, mostramos que a pausa não é uma marca acústica necessária para a segmentação da fala. Temos aqui dois enunciados, como mostrado na Figura 5, na qual o segundo enunciado está destacado. Mesmo sem haver pausa entre eles, podemos interpretá-los em isolamento (áudios 8a e 8b). Contudo, temos diversas pausas dentro do primeiro enunciado: três delas ultrapassam os 600 ms, e uma, em coincidência com o *retracting*, é de 521ms. O *retracting* pode ser considerado um provável engatilhador da pausa, mas constitui um fenômeno separado dela.

Exemplo 9 - bfamcv14 [165-166]

áudio 9

*LCS: ocê descendo a Raja / &he / aí é que eu nũ tem certeza // não / cê não pode descer a Raja direto / e já pegar a Raja indo pra lá //

Figura 6 – Sinal de áudio, espectrograma, F0 e transcrição do exemplo 9



Em 9 também temos duas unidades terminadas. Novamente, não há pausas entre elas, mas duas pausas muito longas dentro da primeira: 1.394ms e 1.085ms. Além disso, entre as pausas, há uma tomada de tempo de 847ms. Não podemos falar de uma única pausa de 3326ms. Os dois fenômenos, pausa e tomada de tempo, devem permanecer separados. Cada um pode aparecer sozinho, como frequentemente acontece. A tomada de tempo, nesse caso, tem provavelmente a função de evitar um silêncio prolongado demais, que geraria um risco alto de perda do turno ou até de encerramento da comunicação.

Exemplos assim não são excepcionais; ao contrário, são comuns. Seja pela função, seja pela frequência, esses casos mostram que a pausa não é um critério adequado para marcar fronteira entre enunciados. Isso é confirmado por uma grande quantidade de casos em que podemos

perceber fronteiras entre unidades autônomas, sem que o fluxo da fala seja interrompido por pausas, como vimos nos exemplos 8 e 9.

2.2 A perspectiva pragmática e a *Language into Act Theory*

Vários autores falam em *pausa virtual*, para se referir a fronteiras entre enunciados não coincidentes com uma interrupção física, mas, sim, com uma pausa "cognitiva". Nespor e Vogel (1986), por exemplo, contrastam o que consideram pausa *virtual* ou pausa perceptual e a pausa real. A pausa perceptual refere-se ao que é percebido como pausa, podendo foneticamente corresponder a uma variedade de fenômenos, como mudança de *pitch* e de duração, e que só algumas vezes corresponde à cessão da fonação. Já a pausa real refere-se à existência de porção de silêncio no sinal da fala. Se, por um lado, isso representa o reconhecimento de que pausa, entendida como silêncio dentro do fluxo da fala, não é um parâmetro adequado para identificar fronteiras entre os enunciados no fluxo da fala, por outro lado, o conceito de *pausa virtual* deixa sem resposta algumas questões: (a) do ponto de vista formal, por que utilizar a expressão *pausa virtual*, que sugere que se trata de uma pausa que não é realmente uma pausa, ou seja, por que denominar um fenômeno de natureza cognitiva, a fronteira, com um termo que normalmente faz referência a algo físico, adicionando um adjetivo (virtual) que indica que o substantivo deve ser entendido como faltante de seu correlato físico?; (b) do ponto de vista mais substancial, o que seria essa *pausa virtual*?; que configuração de parâmetros físicos (*pitch,* duração, intensidade dos fones pré e pós fronteiriços) geram um objeto nomeado de forma que sugere ser apenas virtualmente o que o nome designaria (como pausa sem pausa)?; e (c) do ponto de vista pragmático, ou seja, mais diretamente comunicativo, o que torna um enunciado autônomo e, portanto, suscetível de ser reconhecido como unidade mínima de referência com capacidade comunicativa?

A seguir tentaremos responder às perguntas mediante de conceitos oriundos da *Language into Act Theory*. Para isso, resumiremos a teoria rapidamente. A bibliografia indicada permite maior aprofundamento.

2.2.1 *Language into AcT Theory* (L-AcT)

Partindo de uma perspectiva pragmática, L-AcT define enunciado como "a menor unidade pragmática e prosodicamente autônoma" (CRESTI, 2000). É uma teoria *corpus driven* (CRESTI 2000; MONEGLIA 2011; RASO, 2012c; MONEGLIA; RASO, 2014), fruto de décadas de observação de *corpora* de fala espontânea, que retoma o conceito de ato de fala de Austin (1962) e analisa a fala com base em unidades demarcadas por meio da prosódia. Segundo essa teoria, o que confere autonomia pragmática a um enunciado é o fato de ele veicular uma ilocução, e o que delimita os enunciados é um tipo de fronteira de unidade tonal (CRYSTAL, 1975), realizada acusticamente pelo que denominamos *quebra prosódica terminal*, marcada nas transcrições com a barra dupla (//) (MONEGLIA; CRESTI, 1997).

Pesquisas baseadas em *corpora* de fala espontânea revelaram algumas regularidades:

i. os falantes segmentam naturalmente a fala em unidades pragmática e prosodicamente autônomas: os enunciados. Enunciados contêm uma unidade tonal-informacional nuclear (ilocucionária), necessária e suficiente para veicular a autonomia pragmática;
ii. o enunciado pode ser constituído por mais de uma unidade tonal, que, em princípio, constitui uma unidade informacional, mas apenas uma delas deve necessariamente ser de natureza ilocucionária;
iii. a prosódia funciona como interface entre locução e ilocução, determinando o significado do que é dito: (a) marcando fronteira entre enunciados, ou seja, as unidades de referência da fala maiores do que a palavra; (b) marcando, com as *quebras prosódicas não terminais* (/), fronteiras entre eventuais unidades internas ao enunciado; e c) marcando a função de cada unidade informacional, seja ela de natureza ilocucionária (nesse caso, a prosódia marca também o tipo de ilocução), seja ela de natureza não ilocucionária (marcando a função específica); e

iv. as relações sintáticas estreitas têm como domínio a unidade informacional. As relações entre as unidades informacionais são de natureza semântica e marcadas prosodicamente (CRESTI, 2014).

Uma quebra prosódica terminal marca, portanto, o limite entre uma unidade de referência e outra. As relações informacionais são internas a essa unidade, e as relações sintáticas estreitas são internas às unidades informacionais que compõem a unidade de referência. A unidade de referência articula-se em volta de um núcleo ilocucionário, que pode (mas não necessariamente deve) ser precedido e / ou seguido por unidades não ilocucionárias. O que revela a unidade ilocucionária são principalmente dois fatores: a) a unidade ilocucionária é interpretável em isolamento. É ela que confere autonomia ao enunciado; e b) nenhuma das unidades não ilocucionárias é interpretável em isolamento, somente em conjunto com a unidade ilocucionária.

Esse critério está na base da segmentação dos *corpora* C-ORAL-ROM, C-ORAL-BRASIL e, com algumas variações, do Santa Barbara *Corpus* (DU BOIS *et al.*, 2005), do CorpAfroAs (METTOUCHI; CHANARD, 2010) e do CoSIH (IZRE'EL; HARY; RAHAV, 2001).

2.2.2 Quebra prosódica, unidade tonal e unidade de referência da fala

A noção de *quebra prosódica* nos parece mais adequada do que aquela de *pausa virtual*. Do ponto de vista terminológico, *quebra* remete a uma ruptura que gera fronteiras entre unidades distintas, enquanto *pausa virtual* remete a um efeito cognitivo em ausência de seu veículo físico. Contudo, ainda precisamos definir o conceito de *quebra*. Pesquisas em várias línguas mostram que fronteiras são percebidas de modo constante e homogêneo pelos falantes (HARRIS; UMEDA; BOURNE, 1981; MO; COLE; LEE, 2008; SCHUETZE-COBURN; SHAPLEY; WEBER, 1991; CARLSON; HIRSCHBERG; SWERTS, 2005; SWERTS; COLLIER; TERKEN, 1994). Para verificar a confiabilidade da segmentação de *corpora* com base na percepção de quebras prosódicas é necessário recorrer a testes estatísticos de medição do acordo entre segmentadores. O teste normalmente usado é o Kappa de Fleiss (1971), que mede a probabilidade que o acordo seja devido ou não

ao acaso. Para o C-ORAL-ROM e o C-ORAL-BRASIL o teste foi aplicado com metodologias um pouco diferentes. A experiência do C-ORAL-BRASIL (por ser posterior ao C-ORAL-ROM) testou o acordo entre três segmentadores antes do início da segmentação e depois da primeira revisão das transcrições. A metodologia de aplicação dos testes é descrita detalhadamente em Moneglia *et al*. (2010), Raso e Mittmann (2009) e Mello *et al*. (2012). O teste mostrou um acordo de 0,86 quanto às quebras não terminais e de 0,87 quanto às terminais, e o valor geral de 0,86. Trata-se de um valor de acordo considerado excelente (o valor máximo é 1). Esse resultado mostra a saliência perceptual das quebras prosódicas. Considere-se que os casos de desacordo nunca são devidos à percepção, por um ou mais juízes, de uma quebra avaliada como terminal *versus* a percepção de ausência de quebra por outros juízes. Os desacordos concentram-se na percepção de quebra terminal *versus* percepção de quebra não terminal e de quebra não terminal *versus* ausência de quebra. Isso nos oferece algumas indicações:

i. existe evidência perceptual indiscutível das quebras prosódicas;
ii. existe evidência perceptual indiscutível que permite aos falantes a atribuição do valor terminal ou não terminal da fronteira, dependendo do tipo de quebra prosódica; e
iii. existem quebras mais salientes e menos salientes.

Se interpretadas à luz da L-AcT, as quebras teriam duas funções importantíssimas: (a) a segmentação entre unidades mínimas interpretáveis autonomamente, ou seja, entre unidades de referência da fala, que podemos chamar de *unidades terminadas*; e (b) a eventual segmentação interna dessas em unidades tonais-informacionais, uma das quais necessariamente deve ser ilocucionária.

Mais complexo é definir o conceito de quebra prosódica (em nosso entender, o principal fenômeno que delimita as fronteiras na fala) em termos de seus correlatos físicos. A pausa produz uma fronteira, mas há fronteiras mesmo em ausência de pausas. A literatura aponta vários parâmetros, entre os quais o *reset* da curva de F0 é considerado um dos principais (ao lado do alongamento da sílaba pré-fronteiriça); mas é possível perceber fronteiras mesmo sem evidência de *reset* da F0, por efeito da combinação de outros parâmetros, como forte e repentina

variação na intensidade, na duração e na velocidade de fala (ALBANO LEONI; MATURI, 2002). Entender como se combinam os parâmetros que constituem a quebra prosódica é de grande interesse para o estudo da estruturação da fala, visto a forte relevância perceptual da quebra prosódica e a grande importância da unidade tonal como domínio de muitos fenômenos linguísticos.

Apesar de não sermos ainda capazes de detalhar as combinações de parâmetros que produzem as quebras prosódicas, é indubitável que atribuímos forte relevância à segmentação prosódica do fluxo de fala. Essa segmentação parece ter dois níveis hierarquicamente distintos: (a) a unidade tonal, que empacota a informação segmental em uma unidade prosódica; e (b) a *unidade terminada* que representa uma unidade mínima autônoma no fluxo da fala.

2.2.3 Unidade de referência e ilocução

Os exemplos a seguir, juntamente ao exemplo 8, mostram (i) uma pequena sequência de unidades terminadas compostas por uma única unidade, necessariamente de natureza ilocucionária; (ii) um caso de unidade terminada complexa, formada por várias unidades diferentes em volta de uma unidade ilocucionária; e (iii) um caso (exemplo 8) de unidade terminada formada por mais unidades tonais, entre as quais mais de uma de natureza ilocucionária: a primeira unidade terminada apresenta quatro ilocuções de instrução. Tudo isso nos permite observar o seguinte:

i. a unidade ilocucionária é essencial para realizar uma unidade terminada, ou seja, a unidade terminada não é simplesmente uma cadeia concluída por uma quebra terminal, mas, sim, uma unidade de natureza acional: sem seu núcleo acional ela não se sustenta. Isso é de extrema importância, porque permite entender uma qualidade de natureza pragmática constitutiva da unidade terminada, ou seja, da unidade de referência da fala;
ii. existem unidades terminadas com mais de uma ilocução, nas quais é realizado um sinal prosódico de continuidade depois da produção de cada ilocução, com exceção da última; esse sinal de continuidade sinaliza que, apesar da unidade já possuir um

> núcleo acional, ela não pode ser considerada terminada. Chamamos *stanza* esse tipo de unidade.[6]

Disso, concluímos que a unidade terminada deve possuir duas propriedades: a presença de pelo menos uma ilocução e a marca prosódica de quebra terminal.

Exemplo 10 bfamdl02 [85-87]

áudio 10


*BAL: que cê acha colocar isso aqui //
*BEL: bom // deu certinho //

Exemplo 11 bfammn01 [19]

áudio 11	áudio 11a	áudio 11b	áudio 11c	áudio 11d	áudio 11e
					

Que / na hora que ele lá envinha voltando do [/1] do comércio / que e' foi fazer a compra dele / **a cobra percebeu o cheiro dele** / na hora que ele lá envinha no [/1] no trilho //

O exemplo 10 mostra como enunciados simples são autônomos em virtude da capacidade de veicularem uma força ilocucionária. O exemplo 11 mostra que uma unidade ilocucionária é necessária para a constituição do enunciado (marcada em negrito), mas que é possível termos outras unidades não ilocucionárias que contextualizam a ilocução, sem ser interpretáveis em isolamento. Isso pode ser verificado escutando cada unidade sozinha nos áudios 11a, 11b, 11c, 11d (a única autônoma) e 11e. O exemplo 8 apresenta uma *stanza* com mais unidades ilocucionárias, quase todas, menos a última, acompanhadas por um sinal prosódico de continuidade.

---

[6] Para mais informações sobre a *Stanza*, além da bibliografia sobre a L-AcT já indicada, veja-se Cresti (2010). Veja-se a bibliografia sobre L-AcT para os casos, não mencionados aqui, de ilocuções padronizadas.

3 Relação entre pausas e quebras prosódicas

Até aqui realizamos uma discussão de natureza teórica, ilustrada por diversos exemplos, questionando a utilização da pausa como principal (ou único) parâmetro para identificar fronteiras de enunciados em *corpora* orais. Também discutimos o conceito de unidade de referência da fala e a importância de uma segmentação que represente adequadamente a realidade dos eventos de fala em *corpora* orais.

Nesta seção investigamos sistematicamente os resultados obtidos como consequência da utilização da pausa como critério de segmentação do fluxo da fala em um *corpus* oral. Tal investigação se justifica com o interesse crescente na compilação de *corpora* de fala. Decorre daí a necessidade de aprimorar técnicas de processamento dos dados que permita compilar *corpora* mais rapidamente e com maior fundamentação científica.

A divisão do contínuo da fala em unidades discretas pode ser realizada de forma manual ou automática. A segmentação manual é feita pelo exame do sinal acústico (onda, espectrograma, curva melódica). É um processo demorado, que requer muito treinamento aos segmentadores, além de ser sujeito aos erros naturalmente decorrentes de um processo manual. Na segmentação automática, o áudio é segmentado por uma máquina em unidades definidas operacionalmente. Apresenta como principais vantagens a rapidez e o fato de os erros serem previsíveis e sistemáticos, podendo ser, portanto, reduzidos. Esse sistema de segmentação é amplamente utilizado para a segmentação fonética da fala em unidades menores do que a palavra (SVENDSEN; SOONG, 1987; van HEMERT, 1991; BARBOSA, 2006), mas é ainda muito pouco aplicado para segmentar unidades maiores do que a palavra, ou seja, enunciados e unidades tonais.

A pausa é o parâmetro acústico que mais recebe atenção na identificação de fronteiras no fluxo da fala, sendo comumente associada a fronteiras de enunciados. A simplicidade de identificar interrupções no sinal acústico justifica sua utilização como principal parâmetro na segmentação automática da fala. Entretanto, considerados os exemplos contrários extraídos do *corpus*, nos perguntamos em que proporção as pausas correspondem a fronteiras prosódicas na fala espontânea, e se seria possível relacionar algum valor de duração de pausa a um

determinado tipo de fronteira prosódica (de enunciados ou unidades tonais).

Por meio de uma investigação sistemática do *corpus* C-ORAL-BRASIL, segmentado manualmente em enunciados e unidades tonais, procuramos identificar a coincidência entre pausas e fronteiras de enunciado; e / ou entre pausa e fronteira de unidade tonal interna ao enunciado. Adicionalmente verificamos se há algum valor de duração de pausa que coincida o máximo possível com as fronteiras de enunciados e o mínimo possível com fronteiras de unidades tonais. Desse modo, poderíamos produzir uma segmentação automática baseada em pausas que, pelo menos, reduzisse o trabalho dos segmentadores.

3.1 A fonte para a coleta dos dados

Os dados da pesquisa foram coletados no C-ORAL-BRASIL (RASO; MELLO, 2012). Para uma descrição detalhada da metodologia do *corpus*, veja-se Raso (2012b), Raso e Mello (2014) e Mello (2014). Para uma descrição da metodologia de compilação do C-ORAL-ROM, veja-se Moneglia (2005). Aqui resumimos as principais características do *corpus*, mostrando a natureza dos dados que embasam a pesquisa.

O C-ORAL-BRASIL representa a fala espontânea, definida como a fala planejada ao mesmo tempo em que é executada (NENCIONI, 1983). A arquitetura representa a variedade diatópica da área metropolitana de Belo Horizonte e privilegia a variação diafásica, com atenção também à diastratia. O *corpus* é de fala informal, com cerca de 210.000 palavras. Os textos têm, em média, 1.500 palavras, com poucos textos significativamente menores ou maiores. As situações dividem-se em 75% de contexto privado / familiar, 25% de contexto público. Cada contexto é dividido igualmente em monólogos, diálogos e conversações. A variação diafásica se completa com a maior variedade situacional possível. A razão para privilegiar a variação diafásica é evidente: em uma concepção da fala como comportamento verbalizado, o tipo de ação realizada é decisiva para a sua variação e estruturação. Portanto, a realização de ilocuções variadas é função da variedade dos comportamentos dos falantes e, consequentemente, da variedade das situações comunicativas.

A segmentação é feita em turnos e unidades terminadas.[7] Nessas, há ainda a segmentação em unidades tonais, *retractings* e unidades interrompidas. O texto e o sinal sonoro estão alinhados pelo *software* Winpitch (MARTIN, 2015), e a unidade de alinhamento corresponde ao que é considerada como unidade de referência da fala.

3.2 Metodologia

A amostra utilizada na pesquisa foi composta por trechos abrangendo nove quebras terminais de um mesmo falante, ignorando casos em que a quebra coincidia com fronteira de turno. Os trechos foram extraídos de cada um dos 139 textos do *corpus*, utilizando-se a técnica de amostragem aleatória simples sem reposição. A amostra analisada totalizou 1.251 fronteiras terminais e 2.580 fronteiras não terminais.

Para operacionalizar a análise, definiu-se como "silêncio" qualquer período de interrupção do sinal não provocado pela existência de consoantes desvozeadas, e como "pausa", qualquer duração convencionalizada de silêncio que pudesse ser utilizada para marcação de fronteiras em um modelo de segmentação automática da fala. Estabelecemos que silêncios de duração mínima de 10ms já seriam considerados pausas, definindo, a partir desse valor mínimo, intervalos de 10 em 10 ms, até chegar a 200 ms. O limite de 200 ms foi estabelecido porque, de modo geral na literatura, um silêncio de 200 ms é sempre considerado pausa, e frequentemente intervalos menores também são considerados pausas.

Com isso, convencionalizamos como pausas intervalos de silêncio maiores ou iguais a 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, 120, 130, 140, 150, 160, 170, 180, 190 e 200 ms. Cada pausa encontrada na amostra foi medida e contabilizada de acordo com esses intervalos (uma pausa de 35 ms, por exemplo, é igual ou maior do que 10, 20 e 30 ms). Assim, podemos saber quantas fronteiras seriam identificadas na segmentação automática assumindo-se diferentes valores de duração das pausas.

---

[7] As unidades terminadas podem ser enunciados (possuem um único núcleo ilocucionário) ou *stanzas* (possuem mais de um núcleo ilocucionário).

O objetivo principal era verificar, para cada intervalo, quantas pausas coincidem com fronteiras julgadas por anotadores humanos como marcadas por quebras terminais e quantas delas coincidem com fronteiras marcadas por quebras não terminais. Adicionalmente, quantas fronteiras não coincidem com qualquer pausa conforme convencionalizada.

Para a análise acústica, foi utilizado o *software* Praat (BOERSMA; WEENINK, 2014); os dados foram anotados em planilhas.

Considerando-se um modelo de segmentação automática da fala com base em pausas, a ideia é verificar se existe um valor de duração mínimo de pausa que seria consistentemente associado à fronteira de tipo terminal. Assim, por exemplo, um silêncio observado no sinal de áudio que tivesse duração de 55ms seria utilizado para assinalar fronteira de enunciado caso a pausa fosse estabelecida tanto em 10ms quanto em 20ms, 30ms, 40ms ou 50ms. Já um silêncio observado de 12ms somente seria associado a fronteiras caso a pausa fosse estabelecida como sendo de 10ms.

Com isso, torna-se possível aferir qual intervalo mínimo apresenta a maior coincidência com as fronteiras de tipo terminal, já que sabemos que não é produtivo simplesmente realizar uma equivalência entre qualquer pausa e fronteiras de tipo terminal, como já demonstrado nos exemplos. Imaginando que a pausa pudesse ser considerada uma marca confiável de fronteira de enunciado, é necessário encontrar um determinado valor de duração de pausa que coincida, de maneira estatisticamente significativa, com as fronteiras de enunciados e, ao mesmo tempo, não coincida com as fronteiras de unidades internas ao enunciado (não terminais). E mesmo se aceitarmos uma taxa de coincidência não significativa estatisticamente, a metodologia de segmentação por pausa seria o mais eficaz possível se as pausas coincidissem o máximo possível com as fronteiras terminais, e o mínimo possível com as fronteiras não terminais.

Os percentuais de coincidência entre pausas e fronteiras terminais ou não terminais em cada ponto foram utilizadas para gerar dois modelos de regressão não linear (GRÁFICO 3), os quais atingiram coeficientes de predição de 0,99.

3.3 Resultados

Na Tabela 1, apresenta-se o cômputo do número de fronteiras terminais e não terminais que coincidiram com pausas de valor igual ou maior a cada duração de pausa delimitada. Por exemplo, na décima linha do quadro, identifica-se que foram detectadas 810 fronteiras terminais e 660 fronteiras não terminais coincidentes a silêncio de tamanho igual ou maior a 100ms. Isso significa que a quantidade de coincidências com fronteiras, a cada linha, é inclusiva (conforme explicado na seção anterior), já que cada valor de pausa é compreendido como a duração mínima de silêncio requerida para que um sistema de segmentação automática o identifique como uma fronteira de tipo terminal.

O que vemos é que, à medida que decresce a duração mínima de silêncio, tomado como referência para identificar uma pausa, aumenta a quantidade de fronteiras terminais coincidentes com pausas, mas, ao mesmo tempo, aumenta também a quantidade de fronteiras não terminais coincidentes com a pausa. Em um sistema de segmentação automático baseado em pausas, isso significaria um aumento do número de fronteiras não terminais erroneamente consideradas como fronteiras terminais. Por exemplo, se consideramos o maior valor de pausa (200ms) como a duração mínima de silêncio necessária para identificação de fronteira terminal, temos 746 coincidências com as fronteiras terminais e outros 587 casos de atribuição de fronteira terminal a fronteiras que, perceptualmente, foram julgadas como não terminais.

Tabela 1 – Número de fronteiras terminais e não terminais coincidentes com pausas

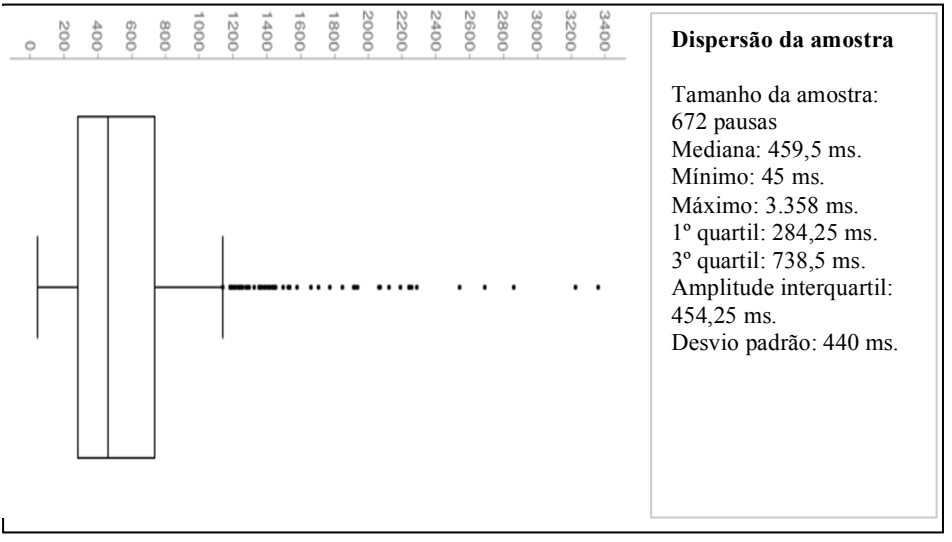
Duração da pausa	Coincidência terminais	Coincidência não- terminais
>= 10 ms	842	675
>= 20 ms	842	675
>= 30 ms	841	675
>= 40 ms	839	675
>= 50 ms	837	673
>= 60 ms	833	673
>= 70 ms	827	668
>= 80 ms	823	667
>= 90 ms	815	663
>= 100 ms	810	660
>= 110 ms	807	655
>= 120 ms	799	646
>= 130 ms	792	639
>= 140 ms	785	630
>= 150 ms	776	629
>= 160 ms	767	623
>= 170 ms	763	615
>= 180 ms	759	607
>= 190 ms	750	595
>= 200 ms	746	587

Se tomamos como parâmetro de segmentação a menor duração de pausa (10ms) como sendo o valor mínimo necessário para identificar uma fronteira terminal, conseguimos identificar no processo automático mais fronteiras terminais (842), mas, ao mesmo tempo, aumentamos também o número de fronteiras não terminais às quais seria erroneamente atribuído o estatuto de terminal (675). Assim, se por um lado temos um benefício de 96 reconhecimentos de fronteiras terminais a mais, por outro lado, temos um aumento de erros com relação às fronteiras não terminais

de 88 fronteiras. Isso sem considerar que a duração de silêncio de 10ms, aquela em que se obtém o melhor resultado na marcação das fronteiras terminais, é normalmente considerado insuficiente para a pausa. Ademais, mesmo aceitando um silêncio de duração tão pequeno como pausa, continuaríamos não capturando 33% das fronteiras que foram marcadas como terminais no C-ORAL-BRASIL, com uma grande consistência entre os anotadores, pois essas não coincidem com pausas convencionalizadas.

Das 1.251 fronteiras terminais analisadas, 409 não coincidem com nenhum valor de silêncio, o que equivale a 33%. Do total de 1.627 enunciados analisados, 913 enunciados apresentam fronteiras não conclusivas em seu interior, dos quais 565 (62%) não apresentam pausas.

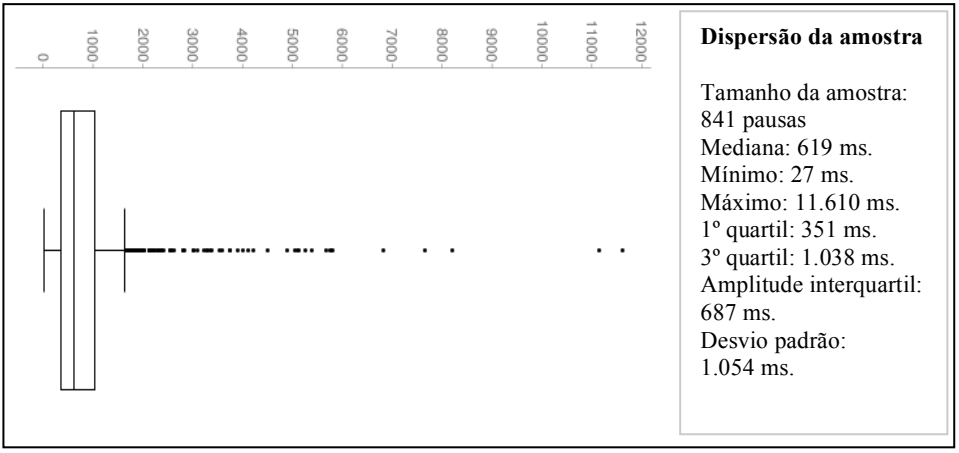
Gráfico 1 – Distribuição das durações de pausas coincidentes com fronteiras não terminais



O Gráfico 1 mostra as durações de pausas da amostra coincidentes com fronteiras identificadas como não terminais. Verifica-se, pelos valores de dispersão da amostra, que a menor duração de silêncio coincidente com fronteira não terminal foi de 45ms, e o valor máximo de silêncio coincidente com fronteira equivalente a 3.229 ms. Podemos

observar que há grande quantidade de pausas não coincidentes com fronteira terminal com duração de silêncio superior a 200ms.

Gráfico 2 – Distribuição das durações de pausas coincidentes com fronteiras terminais



O Gráfico 2 traz as durações de pausas coincidentes com fronteiras identificadas pelos transcritores do *corpus* como terminais. Nota-se que as pausas coincidentes com fronteiras terminais apresentam uma dispersão ainda maior do que aquelas coincidentes com fronteiras não terminais. Os valores de dispersão mostram que o valor mínimo de pausa coincidente com fronteira terminal foi de 27 ms, e o valor máximo equivalente a 11.610 ms. Comparando-se os dados dos Gráficos 1 e 2, podemos observar a grande sobreposição em termos das durações de pausas coincidentes com fronteiras não terminais e terminais. Metade (50%) das pausas tem durações entre 351 ms e 1.038 ms, o que engloba a maioria dos valores de pausa coincidentes com fronteiras não terminais. O desvio padrão é de 1.054 ms (praticamente 1 s de duração). Nota-se que as pausas na fala espontânea são distribuídas de modo bastante disperso, sendo, portanto, muito difícil correlacionar a duração com a marcação de um tipo particular de fronteira.

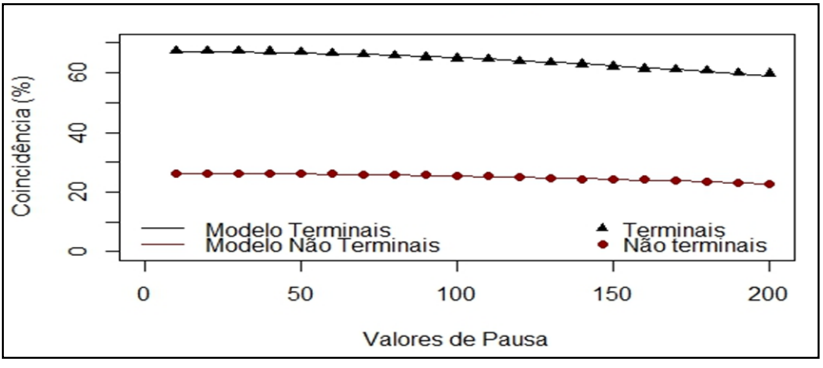
Para averiguar se há algum valor de duração de silêncio que possa ser utilizado como valor mínimo para marcação de fronteiras de

enunciados na fala espontânea, foram criados dois modelos não lineares que indicam a relação entre os diferentes valores mínimos convencionalizados para pausa e a taxa de coincidência com fronteiras de tipo terminal e não terminal.

As curvas expostas no Gráfico 3 descrevem a relação entre o tamanho convencionalizado da pausa e a coincidência entre pausa e fronteira terminal. Os ícones indicam os valores observados e as linhas contínuas, os modelos aproximados. Os modelos estimados (linhas contínuas) para o Gráfico 3 são representados pelas equações seguintes:

- para as fronteiras terminais: $y = 67{,}19 * \exp(-0{,}00000327*x^2)$
- para as fronteiras não terminais: $y = 26{,}33 * \exp(-0{,}00000353*x^2)$

Gráfico 3 – Percentual de coincidência entre pausas e fronteiras terminais e não-terminais, por duração mínimo de pausa



No Gráfico 3, a curva superior indica a relação entre pausas de diferentes durações e fronteiras terminais do *corpus*. Na vertical, temos as porcentagens de coincidência entre pausas e fronteiras. Nota-se que quando o valor da pausa tende a 0ms, há uma coincidência limite máxima de 67,2% entre fronteira terminal e pausa. Para a curva das fronteiras não terminais esse limite é de 26%. A partir daí, as duas curvas decaem.

Analisando o modelo para as fronteiras não terminais, observamos que a coincidência entre pausa e fronteira chega ao valor mínimo de 19,2% quando o valor mínimo de duração de pausa é de 300 ms ou mais.

A coincidência entre pausas e fronteiras não-terminais chega abaixo dos 5% somente com pausas com duração de, pelo menos, 687 ms, resultado consistente com a distribuição das durações de pausas mostradas no Gráfico 1, no qual observamos que 75% das pausas têm duração de até 738,5 ms.

Ainda no Gráfico 3, para o valor de pausa maior ou igual a 100ms, observa-se 65% de coincidência com as fronteiras terminais e 25,4% de coincidência com as fronteiras não-terminais no *corpus*. A segmentação, segundo esse critério, deixaria de identificar, portanto, 35% das fronteiras terminais existentes e marcaria como terminais 25,41% das fronteiras que, na nossa segmentação, são não terminais.

Verifica-se, então, que se variando o tamanho de pausa, as duas curvas apresentam comportamentos semelhantes: ao aumentar o tamanho de pausa, aumentam as coincidências nos dois casos; ao diminuir o tamanho de pausa, ocorre o inverso. Logo, não é possível estabelecer uma duração de pausa que tenha a máxima coincidência com fronteiras terminais e a mínima coincidência com fronteiras não terminais.

4 Conclusão

Os dados expostos indicam que

i. o critério de segmentação por pausa não se mostra adequado para a segmentação de textos de fala espontânea. Não somente os exemplos mostrados nesse trabalho, mas também os modelos estatísticos realizados sobre amostra aleatória extraída do *corpus* mostram que não há uma real correspondência entre pausas e fronteiras de enunciados e unidades menores (já validadas quanto ao acordo entre anotadores); a pausa não parece constituir a única (e talvez nem seja a principal) estratégia de marcação de fronteiras na fala espontânea; e

ii. não existe uma duração de pausa que possa ser eficazmente empregada para marcação de fronteira terminal de unidade de referência. O aumento na captura de fronteiras terminais, dada uma certa duração de pausa, corresponde também a superestimativa de fronteiras, e o aumento no reconhecimento de fronteiras não terminais corresponde a subestimativa de fronteiras.

As conclusões desta pesquisa são, a nosso ver, muito claras: qualquer definição de pausa relacionada a um correlato físico de silêncio não pode ser automaticamente considerada marca de fronteira de unidade de referência da fala. O conceito de quebra prosódica é mais eficiente para essa função demarcativa tão importante. Contudo, a questão da segmentação da fala não se conclui aqui. Precisamos entender melhor como se realiza a quebra prosódica em termos de seus correlatos físicos. Sabemos que as quebras possuem uma fortíssima saliência perceptual, mas não sabemos exatamente de que maneira atuam os correlatos físicos que as produzem.

Nesse sentido, um desdobramento natural, já encaminhado do trabalho apresentado aqui, é o desenvolvimento de um modelo computacional que considere diversos parâmetros acústicos utilizados na identificação de fronteiras. Com uma abordagem *corpus driven* e o conceito de aprendizagem de máquina, visa-se compreender melhor as configurações físicas das quebras, mediante uma amostra de fala espontânea segmentada em enunciados e unidades tonais por múltiplos juízes, da qual seriam extraídos dados relativos a diversos parâmetros acústicos (por exemplo, F0 média de sílaba pré e pós fronteiriça e duração silábica). Com base nos resultados obtidos será possível delinear as características prosódicas das fronteiras e calcular sua probabilidade de ocorrência.

Acreditamos que uma resposta, mesmo que parcial, à pergunta acerca das propriedades físicas constitutivas das fronteiras teria dois efeitos fundamentais para o estudo da fala: o primeiro seria permitir uma segmentação (semi)automática de *corpora* de fala. Para quem sabe quão grande é o trabalho de compilação de *corpora* de fala, é fácil imaginar a vantagem que isso traria: seria muito mais viável compilar *corpora*, não de poucas centenas de milhares de palavras, mas de milhões de palavras,

com os óbvios reflexos disso para a pesquisa em diferentes áreas. O segundo efeito seria de natureza teórica: conhecer mais sobre os correlatos físicos das fronteiras entre as unidades de referência da fala abriria perspectivas importantes para compreender melhor a percepção da fala e, por consequência, seus mecanismos neurolinguísticos.

Referências

ALBANO LEONI, F.; MATURI, P. *Manuale di fonetica*. Roma: Carocci, 2002.

AUSTIN, J. *How to do things with words*. London: Oxford University Press, 1962.

BARBOSA, P. A. *Incursões em torno do ritmo da fala*. São Paulo: Pontes, 2006. (Fapesp).

BERRUTO, G. Varietà diamesiche, diastratiche, diafasiche. In: A SOBRERO, A (Org.). *Introduzione all'italiano contemporaneo***:** La variazione e gli usi. Roma: Laterza, 1993. p. 37-92.

BIBER, D. *et al*. *The Longman grammar of spoken and written English***.** London: Longman, 1999.

BOERSMA, P.; WEENINK, D. *Praat*: doing phonetics by computer. 2014. Programa de computador, versão 5.4.19. Disponível em: <http://www.praat.org/>. Acesso em: 16 set. 2015.

BUHMANN, Jeska *et al*. Annotation of prominent words, prosodic boundaries and segmental lengthening by non-expert transcribers in the Spoken Dutch Corpus. In: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION, 3, 2002, Las Palmas. *Proceedings...* Las Palmas: Elra, Universidad de Las Palmas de Gran Canaria, 2002. p. 1 - 7. Disponível em: <http://www.lrec-conf.org/proceedings/lrec2002/pdf/96.pdf>. Acesso em: 20 set. 2015.

CARLSON, R.; HIRSCHBERG, J.; SWERTS, M. Cues to upcoming Swedish prosodic boundaries: Subjective judgment studies and acoustic correlates. *Speech Communication***,** Nara, v. 46, n. 3-4, p. 326-333, jul.

2005. Elsevier BV. Disponível em: <http://api.elsevier.com/content/article/PII:S0167639305000932?httpAccept=text/xml>. Acesso em: 22 set. 2015.
DOI: <http://dx.doi.org/10.1016/j.specom.2005.02.013>

CAVALCANTE, F.; RAMOS, A. Um *minicorpus* de inglês americano etiquetado informacionalmente, em preparação.

CHOMSKY, N. Remarks on nominalization. In: JACOBS, R.; ROSENBAUM, P. (Org.) *Reading in English Transformational Grammar*. Waltham: Ginn, 1970. p. 184-221.

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

CRESTI, E., La Stanza: un'unità di costruzione testuale del parlato. In: CONGRESSO DELLA SOCIETÀ INTERNAZIONALE DI LINGUISTICA E FILOLOGIA ITALIANA, 10, 2008, Basilea. *Atti*… Org.: A. Ferrari. Firenze: Cesati, 2010. p. 713 - 732.

CRESTI, E. Syntactic properties of spontaneous speech in the Language into Act Theory. In: RASO, T.; MELLO, H. *Spoken Corpora and Linguistic Studies*. Amsterdam/Philadelphia: John Benjamins, 2014. p. 365-410.
DOI: <http://dx.doi.org/10.1075/scl.61.13cre>

CRESTI, E.; GRAMIGNI, P. Per una linguistica corpus based dell'italiano parlato: Le unità di riferimento. In: CONVEGNO NAZIONALE "IL PARLATO ITALIANO", Napoli. *Atti*… Ed.: F. Albano Leoni, *et al*. Napoli: D'Auria, CD-ROM, 2004. p. 1–26.

CRESTI, E.; MONEGLIA, M. (Orgs.). *C-ORAL-ROM*: Integrated reference corpora for spoken Romance Language*s*. Amsterdam/Philadelphia: John Benjamins, 2005.
DOI: <http://dx.doi.org/10.1075/scl.15>

CRYSTAL, D. *The English tone of voice*. London: Edward Arnold, 1975.

DU BOIS, J. *et al*. *Santa Barbara Corpus of Spoken American English,* Parts 1-4. Philadelphia, PA: Linguistic Data Consortium, 2000–2005.

FLEISS, J. L. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, v. 76, p. 378-382, 1971.
DOI: <http://dx.doi.org/10.1037/h0031619>

HARRIS, Z. S. *String analysis of sentence structure***.** The Hague: Mouton, 1962.

HARRIS, M.; UMEDA, N; BOURNE, J. Boundary perception in fluent speech. *Journal of Phonetics*. n. 9, p.1-18, 1981.

IZRE'EL, S.; HARY, B.; RAHAV, G. Designing CoSIH: The Corpus of Spoken Israeli Hebrew. *International Journal of Corpus Linguistics*, n. 6, p. 171-197, 2001.

MARTIN, Philippe. Winpitch Pro W8. v 7.2.00. Pitch Instruments. Disponível em: <http://www.winpitch.com/>, 2015. Acesso em: 25 set. 2015.

MARUYAMA, T. Speech segmentation by clausal and non-clausal boundaries in Japanese. INTERNATIONAL CONFERENCE ON COGNITIVE SCIENCE, 5, Kaliningrad. *Abstracts… V. 2*: Workshop Spoken discourse corpora as a window on cognitive mechanisms of speech production. Kaliningrad, June 2012. p. 787-783

MARUYAMA, T. Two-Level Utterance Units: Cognitive and Communicative Aspects of Spontaneous Speech. LABLITA, 9; LEEL, 4, INTERNATIONAL WORKSHOP, Belo Horizonte, 2015. *Conferências*... Units of Reference for Spontaneous Speech Analysis and their correlations across languages, Belo Horizonte, August 6, 2015.

MELLO, H. R., Methodological issues for spontaneous speech corpora compilation: The case of C-ORAL-BRASIL. In: RASO, T.; MELLO, H. R. *Spoken Corpora and Linguistic Studies.* Amsterdam/Philadelphia: John Benjamins, 2014. p. 27-68.
DOI: <http://dx.doi.org/10.1075/scl.61.01mel>.

MELLO, H. R. *et al*. Transcrição e segmentação prosódica do *corpus* C-ORAL-BRASIL: critérios de implementação e validação. In: RASO, T.; MELLO, H. R. (Ed.) *C-ORAL – Brasil I*: *Corpus* de referência do português brasileiro falado informal. Belo Horizonte: Editora UFMG, 2012. p. 125-176.

MELLO, H.; RASO, T.; BOSSAGLIA, G.; SANTANA, T. Enunciados e estruturas de predicação no *corpus* C-ORAL-BRASIL. (Em preparação)

MERLO, S.; BARBOSA, P. A. Séries temporais de pausas e de hesitações na fala espontânea. *Caderno de Estudos Linguísticos.* v. 1, n. 54, p. 11-24, 2012. Disponível em: <http://revistas.iel.unicamp.br/index.php/cel/article/view/2569>. Acesso em: 20 set. 2015.

METTOUCHI, A.; CHANARD, C. From Fieldwork to Annotated Corpora: the CorpAfroAs Project. *Faits de Langue-Les Cahiers.* n. 2, p. 255-265, 2010.

MITTMANN, M. M. O C-ORAL-BRASIL e o estudo da fala informal: um novo olhar sobre o Tópico no Português Brasileiro. Tese - (Doutorado em Estudos Linguísticos) Faculdade de Letras da Universidade Federal de Minas Gerais, UFMG, Belo Horizonte, 2012, 248 p., Belo Horizonte, 2012. 248 p. Disponível em: <http://www.bibliotecadigital.ufmg.br/dspace/ handle/1843/LETR-97YMKT>. Acesso em: 20 set. 2015.

MITTMANN, M. M. Análise da estruturação de diálogos e monólogos na fala informal: quantificando as diferenças, *Domínios de Lingu@gem*, v. 7, p. 338-372, 2013.

MITTMANN, M.; RASO, T. The C-ORAL-BRASIL Informationally Tagged Mini-Corpus. In: MELLO, H. R.; PANUNZI, A.; RASO, T. (Org.). Illocution, modality, attitude, information patterning and speech annotation. Firenze: Firenze University Press, 2012. p. 151-183.

MO, Y.; COLE, J.; LEE, E. Naïve listeners' prominence and boundary perception. In: SPEECH PROSODY INTERNATIONAL CONFERENCE, 4, 2008, Campinas. *Proceedings...* Org.: P. A. BARBOSA; S. MADUREIRA; C. REIS (Org.). Campinas: Isca, 2008. p. 735 - 738. Disponível em: <http://www.isca-speech.org/archive/sp2008/sp08_735.html>. Acesso em: 20 set. 2015.

MONEGLIA, M. The C-ORAL-ROM resource. In: CRESTI, E.; MONEGLIA, M. (Org.). *C-ORAL-ROM*: integrated reference corpora for spoken Romance languages. Amsterdam/Philadelphia: John Benjamins, 2005. p. 1-70. DOI: <http://dx.doi.org/10.1075/scl.15.03mon>

MONEGLIA, M. Spoken corpora and pragmatics. *Revista Brasileira de Linguística Aplicada*, Belo Horizonte, v. 11, n. 2, p. 479-519, 2011. Disponível em: <http://www.periodicos.letras.ufmg.br/rbla/arquivos/335.pdf>.

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

MONEGLIA, M. *et al.* Evaluation of consensus on the annotation of terminal and non-terminal prosodic breaks in the C-ORAL-ROM corpus. In: CRESTI, E.; MONEGLIA, M. (Org.). *C-ORAL-ROM*: integrated reference corpora for spoken Romance languages. Amsterdam: John Benjamins, 2005. p. 257-276.

MONEGLIA, M.; RASO, T. Notes on Language into Act Theory. In: RASO, T; MELLO, H. R. *Spoken Corpora and Linguistic Studies*. Amsterdam / Philadelphia: John Benjamins, 2014. p. 468-495. DOI: DOI: <10.1075/scl.61.15mon>

MONEGLIA, M. *et al*. Challenging the perceptual relevance of prosodic breaks in multilingual spontaneous speech corpora: C-ORAL-BRASIL/C-ORAL-ROM. In: SPEECH PROSODY, Chicago. *Proceedings*... Chicago, 2010. p. 1-4.

NENCIONI, G. *Di scritto e di parlato*: discorsi linguistici. Bologna: Zanichelli, 1983.

NESPOR, M.; VOGEL, I. *Prosodic phonology.* Dordrecht: Foris, 1986.

PANUNZI, A.; MITTMANN, M. The IPIC resource and a cross-linguistic analysis of information structure in Italian and Brazilian Portuguese. In: RASO, T.; MELLO, H. R. Spoken Corpora and Linguistic Studies. Amsterdam / Philadelphia: John Benjamins, 2014. p. 129-151. DOI: <http://dx.doi.org/10.1075/scl.61.05pan>.

RASO, T. Minicorpus retirado do corpus C-ORAL-BRASIL e etiquetado informacionalmente. Annotation of information patterns according to Language into Act Theory in DB IPIC - first release, 2012a. Disponível em: <http://lablita.dit.unifi.it/ipic/>.

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

RASO, T.; O C-ORAL-BRASIL e a Teoria da Língua em Ato. In: RASO, T; MELLO, H. R. (Org.). *C-ORAL-BRASIL I*: *Corpus* de referência do português brasileiro falado informal. Belo Horizonte: Editora UFMG, 2012c. p. 91-124.

RASO, T. Fala e escrita: meio, canal, consequências pragmáticas e linguísticas. *Domínios de Lingu@gem*, v. 7, p. 12-46, 2013.

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

RASO, T.; MITTMANN, M. As principais medidas da fala. In: RASO, T.; MELLO, H. R. (Org.). *C-ORAL – BRASIL I*: *Corpus* de referência do português brasileiro falado informal. Belo Horizonte: Editora UFMG, 2012. p. 177-222.

RASO, T.; MITTMANN, M. Validação estatística dos critérios de segmentação da fala espontânea no *corpus* C-ORAL-BRASIL. *Revista de Estudos da Linguagem*, v. 17, p. 73-92, 2009.
DOI: <http://dx.doi.org/10.17851/2237-2083.17.2.73-91>.

ROSSI, F. La variazione diamesica. In: SIMONE, R. (Org.). *Enciclopedia dell'Italiano*. Milano: Treccani, 2011. Disponível em: <http://www.treccani.it/enciclopedia/variazione-diamesica_(Enciclopedia_dell'Italiano)/>. Acesso em: 20 set. 2015.

SCHUETZE-COBURN, S; SHAPLEY, M; WEBER, E. G. Units of intonation in discourse: a comparison of acoustic and auditory analyses. *Language and speech*. v. 34, n. 3, p. 207-234, 1991. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/1843524.0023830991034>. Acesso em: 20 set. 2015.

SORIANELLO, P. *Prosodia.* Roma: Carocci, 2006.

SVENDSEN, T.; SOONG, F. On the automatic segmentation of speech signals. IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING - ICASSP '87, Dallas, 1987.

Conferences... Dallas, US: Institute of Electrical & Electronics Engineers (IEEE), 1987. p. 77-80. Disponível em: <http://dx.doi.org/10.1109/ICASSP.1987.1169628>. Acesso em: 20 set. 2015.

SWERTS, M.; COLLIER, R.; TERKEN, J. Prosodic predictors of discourse finality in spontaneous monologues. *Speech Communication*, v. 15, n. 1-2, p. 79-90, 1994. Disponível em: <http://www.sciencedirect.com/science/article/pii/0167639394900434>. Acesso em: 20 set. 2015.

VAN HEMERT, J. P. Automatic segmentation of speech. *Ieee Transactions On Signal Processing***,** Piscataway, v. 39, n. 4, Institute of Electrical & Electronics Engineers (IEEE), p. 1008-1012, abr. 1991. DOI: <http://dx.doi.org/10.1109/78.80941>.

ZELLNER, B. Pauses and the temporal structure of speech. In: KELLER, E. (Org.) *Fundamentals of speech synthesis and speech recognition*. Chichester: John Wiley, 1994. p. 41-62.