Validação estatística dos critérios de segmentação da fala espontânea no *corpus* C-ORAL-BRASIL

Statistic validation of speech segmentation criteria in the C-ORAL-BRASIL corpus

Tommaso Raso
Universidade Federal de Minas Gerais - UFMG
Maryualê Malvessi Mittmann
Universidade Federal de Minas Gerais - UFMG

Resumo

Este artigo apresenta o processo empregado no *corpus* C-ORAL-BRASIL, bem como os resultados estatísticos da segmentação da fala espontânea encontrados nesse *corpus*. Exploramos especialmente a validação estatística dos critérios para a segmentação da fala em enunciados e unidades tonais com base na Teoria da Língua em Ato. A validação teve por objetivo assegurar que o produto final tivesse a maior uniformidade possível quanto à segmentação da fala. Após um ano, realizamos novos testes de validação, para reavaliação do grupo principal de transcritores no momento da revisão do *corpus*. Os resultados da validação estatística ao final da fase de treinamento indicam alta confiabilidade quanto à segmentação do *corpus*, e a reavaliação indicou um grau ainda maior de acordo entre os transcritores. O principal objetivo deste processo, a confiabilidade e uniformidade das transcrições, foi obtido.

Palavras-chave

Corpus, Segmentação da fala, Fala espontânea

Abstract

This paper presents the training process and the statistic results for spontaneous speech segmentation employed in the C-ORAL-BRASIL corpus. We explore in detail the statistic validation of segmentation criteria used to parse the speech flow into utterances and tone units according to the Informational Patterning Theory. The aim of the validation was to assure that the final product has the greatest uniformity possible concerning the speech segmentation. A year after the training, we performed new validation tests, to re-evaluate the main group of transcribers at the moment of the corpus revision. The results of the statistic validation that took place by the end of the training point to a high segmentation reliability, and the re-evaluation results show an even higher degree of agreement between the transcribers. The major goal of this process, which was the uniformity and reliability of the transcriptions, was achieved.

Keywords

Corpus, Speech Segmentation, Spontaneous Speech

1. Introdução

O objetivo deste artigo é apresentar os critérios de segmentação da fala em unidades tonais/informacionais, a metodologia de treinamento dos transcritores e a validação estatística das segmentações realizados em um *corpus* de fala espontânea do português do Brasil (PB), o *corpus* C-ORAL-BRASIL. Apresentamos em detalhe o processo de treinamento de transcritores e a validação estatística relativa à semelhança da anotação prosódica do *corpus* entre diferentes transcritores.

O *corpus* C-ORAL-BRASIL é a base de um projeto[1] coordenado por Heliana Mello e Tommaso Raso, sediado na Universidade Federal de Minas Gerais. O objetivo do projeto é o estudo da fala espontânea do PB, comparativamente às principais línguas românicas europeias. O C-ORAL-BRASIL segue os mesmos critérios de arquitetura e segmentação dos corpora do projeto C-ORAL-ROM (italiano, francês, espanhol e português europeu) (CRESTI; MONEGLIA, 2005) e permite, além de outros objetivos, o estudo da fala com base na Teoria da Língua em Ato[2] (CRESTI, 2000; MONEGLIA, 2005; CRESTI, 2008).

2. O *Corpus*

O *corpus* C-ORAL-BRASIL[3] será composto por um total de 300.000 palavras em 200 textos, que se dividem em uma metade informal e uma metade formal. A metade informal, a mais importante para o estudo da fala espontânea, está prestes a ser concluída. A metade formal vai ser implementada em seguida. A parte informal compõe-se de 100 textos e 150.000 palavras, assim organizados: 80% de domínio familiar/privado e 20% de domínio público; em cada domínio, um terço de diálogos, um terço de monólogos e um terço de conversações (interações entre mais de duas pessoas). Cerca de 80% dos textos têm aproximadamente 1.500 palavras; os 20% restantes poderão incluir textos menores, desde que textualmente orgânicos, e também textos de cerca de 4.500 palavras.

O C-ORAL-BRASIL não tem por finalidade buscar um balanceamento em relação às variações diatópica e diatrática. Em relação à diatopia, a maioria dos falantes pertence à região de Minas Gerais, principalmente à área metropolitana de Belo Horizonte. Já quanto à diastratia, todos os níveis socioculturais e faixas etárias são representados, mas de maneira não sistemática; neste quesito apenas a proporção entre homens e mulheres é, praticamente, paritária.

A característica principal do *corpus* é a grande representação da dimensão diafásica da fala, uma vez que contém textos de fala espontânea coletados em diferentes situações comunicativas, em contextos naturais. O objetivo principal da arquitetura do *corpus* é abranger a maior variedade possível de situações comunicativas, para possibilitar o estudo da fala com base nas ações (AUSTIN, 1962) executadas através de cada enunciado da fala e da estrutura informacional desses enunciados, conforme proposto por Cresti (1995; 2000; 2008).

Alguns exemplos de situações presentes no *corpus* são: pessoas trabalhando em obra, jogando um jogo de mesa, jogando futebol, conversando enquanto dirigem, fazendo compras, ensinando e aprendendo a usar um equipamento, interagindo com o professor durante uma aula de ginástica, garçons preparando o necessário para uma festa e interagindo com os hóspedes, uma mãe contando uma história para uma criança, contos de piadas, de receitas, de vida, de trabalho, etc.

Para possibilitar a gravação da fala de duas ou mais pessoas em situações naturais do cotidiano, foi utilizado um equipamento de gravação de alta qualidade com microfones sem fio. Isto permite, ao menos, a extração da curva da frequência fundamental (F0), que é o principal parâmetro prosódico utilizado na análise das ilocuções e da estrutura informacional da fala espontânea. Apesar do ruído inevitável em muitas situações, frequentemente outros dados acústicos também podem ser extraídos com certa qualidade.

Os textos foram ou estão sendo transcritos, revisados e alinhados para a análise, esta última etapa realizada através do software WinPitch, de Ph. Martin (MARTIN, 2004) exclusivamente por transcritores *experts*, formados previamente para isso. Um *subcorpus* de 20 textos e pouco mais de 30.000 palavras, extraído respeitando-se proporcionalmente a arquitetura do *corpus*, está sendo etiquetado informacionalmente.

3. Critérios de segmentação da fala segundo a Teoria da Língua em Ato

A Teoria da Língua em Ato analisa a fala como sequencia de enunciados. O enunciado é entendido como a "contraparte linguística do ato de fala", ou seja, corresponde, no ato de fala, ao ato locutório, é a menor unidade da fala interpretável pragmaticamente (CRESTI, 2000). A fronteira entre enunciados é marcada por uma quebra entonacional percebida pelos falantes como terminal (conclusiva). O perfil conclusivo ou terminal independe do tipo de ilocução veiculada pelo enunciado. Cresti (2000) também classifica o enunciado em simples ou complexo. O enunciado é simples se formado por uma única unidade tonal (a unidade de comentário, que tem a função de veicular a força ilocucionária e é, portanto, necessária). É complexo quando formado pela unidade de comentário e uma ou mais unidades tonais com diferentes funções informacionais. As unidades tonais/informacionais dentro de um enunciado são separadas por uma quebra entonacional percebida como não terminal (não conclusiva).

Para representar as quebras prosódicas no formato de transcrição da fala adotado no C-ORAL-BRASIL, utiliza-se a seguinte notação:

// - barra dupla – representa quebra de valor terminal e marca a conclusão de enunciado.

/ - barra simples – representa quebra de valor não terminal e marca a fronteira de unidade tonal.

+ - marca a interrupção de enunciados, independentemente do motivo pelo qual eles são abandonados antes da sua conclusão.

[/nº] - marca os fenômenos de *retracting*; o número que segue a barra indica o número de palavras canceladas na retomada do texto.[4]

Tanto a barra dupla quanto o símbolo "+" assinalam, portanto, quebras de natureza terminal, enquanto a barra simples e a barra entre colchetes representam quebras de natureza não terminal.

O valor informacional de cada unidade é definido com base em três critérios: a função, o perfil entonacional (cada unidade é caracterizada por perfis entonacionais dedicados a certas funções) e a distribuição relativa à unidade de comentário. A entonação tem o papel de atuar como a interface entre os atos locutório e ilocutório. Ela realiza a segmentação da fala em unidades mínimas

interpretáveis pragmaticamente (os enunciados) e em unidades melódicas internas ao próprio enunciado (CRESTI, 2000).

O mesmo conteúdo locutivo pode veicular ilocuções diferentes, dependendo de como é entoado, como também pode veicular estruturas informacionais diferentes. Vejamos alguns exemplos simples de algumas das possibilidades em que o mesmo conteúdo locutivo "João vai pro Rio" é expresso em diferentes padrões tonais/informacionais:[5]

(1) **João vai pro Rio //**

(2) João **/ vai pro Rio //**

(3) **João /** vai pro Rio **//**

Esse conteúdo locutivo pode ser entoado com diversos valores ilocucionários: como uma resposta, uma pergunta total, uma ordem, uma expressão de surpresa, etc. Além disso, pode também ser informacionalmente estruturado de maneiras diferentes: (1) seria um enunciado simples, composto unicamente pela unidade de comentário; já (2), dependendo da maneira como as unidades são entoadas, pode representar uma unidade informacional de tópico (que determina o âmbito de aplicação da força ilocucionária) e uma unidade de comentário, ou ainda uma unidade de alocutivo (para dirigir-se a "João") e uma de comentário. O exemplo (3) representa a veiculação de unidade de comentário seguida por uma unidade de apêndice. Cada ilocução e cada estruturação informacional corresponde a um padrão prosódico diferente e característico de um certo valor ilocucionário ou de um padrão informacional. O que é comum a cada ilocução é a forma e posição do núcleo, independente da parte preparatória, ou conclusiva, ligada a um diferente tamanho (silábico) da locução (FIRENZUOLI, 2003).

4. A validação estatística da segmentação da fala

Os critérios para a segmentação da fala em enunciados e unidades tonais são de natureza perceptual. Os transcritores são orientados a se guiarem exclusivamente por sua percepção acerca da presença ou não de quebras e do caráter terminal ou não terminal das quebras prosódicas executadas pelo informante. Esse método de trabalho gera um certo nível de irregularidade no produto final, visto que diferentes pessoas podem perceber (e, portanto, segmentar) de maneira diferente quebras perceptualmente mais fracas, dependendo de fatores que vão da variedade linguística do transcritor ao seu nível de atenção.

Dado que a percepção apresenta um certo grau de diferença entre os indivíduos, sentiu-se a necessidade de elaborar uma metodologia de treinamento para os futuros transcritores que garantisse a maior uniformidade possível na segmentação das transcrições, independentemente do fato de terem sido executadas por transcritores diferentes. A formação como um todo envolveu uma fase de formação teórico-prática e sessões de treinamento e teste de acordo entre os transcritores. O acordo foi avaliado pelo teste Kappa[6] (FLEISS, 1971).

4.1. Metodologia de formação para a transcrição

Considerando a experiência do grupo do C-ORAL-ROM, verificou-se que a formação melhorou o desempenho dos transcritores do C-ORAL-BRASIL se comparado com o projeto europeu. A formação envolveu fases de treinamento e discussões relativas tanto à transcrição do nível segmental da fala quanto à percepção das fronteiras de enunciados e unidades tonais (segmentação prosódica). Aqui serão expostos somente os aspectos da formação relativos à segmentação prosódica. Contudo, é necessário ressaltar que o nível segmental foi transcrito com base ortográfica, mas que em muitos pontos foram estabelecidos critérios não ortográficos. Isso teve a finalidade de preservar importantes fenômenos morfossintáticos e lexicais da fala que inevitavelmente se perderiam com uma transcrição plenamente ortográfica.[7] Esse aspecto é relevante para os nossos fins neste artigo, uma vez que a transcrição do nível segmental e a marcação em nível prosódico acontecem juntas, sendo ambas atividades baseadas na percepção.

A formação para segmentação prosódica, sinteticamente, realizou-se da seguinte maneira. Primeiramente, os transcritores potenciais passaram por uma etapa de formação acadêmica com o propósito de instrumentalizá-los na base teórica e nos critérios de segmentação. Essa formação foi realizada durante um curso de pós-graduação de 60 horas; quatro minicursos, dois de 15 horas e dois de 8 horas cada um; bem como três *workshops* de 8 horas cada um.[8] Depois desta primeira fase, dez alunos de doutorado, mestrado ou iniciação científica foram escolhidos para um treinamento mais específico. O treinamento foi planejado em conjunto com os responsáveis pelo projeto europeu, M. Moneglia e E. Cresti, e coordenado por Tommaso Raso. Esta etapa consistiu em um processo de treinamento e *feedback*, no qual um grupo de transcrições já prontas foi utilizado para testar o acordo entre transcritores quanto à segmentação prosódica dos textos. Essas transcrições, escolhidas por serem consideradas de

dificuldade média, eram fruto de exercitações feitas durante as atividades realizadas antes. Isso significa que as transcrições apresentavam vários defeitos, o que, como se verá, deve ser considerado na interpretação dos resultados dos testes de acordo entre transcritores. Os defeitos eram, por exemplo, a falta de transcrição de um enunciado ou de parte dele, ou transcrição errada de pequenos trechos. Ainda assim, julgamos que, para a realização dos testes, não era necessário revisar as transcrições. Naturalmente, antes dos testes eliminou-se das transcrições toda a anotação da segmentação prosódica.

Para os testes, os transcritores foram divididos em três grupos, com base nos seguintes critérios: B (na época aluna de IC e logo depois mestranda), M (aluna de doutorado) e H (mestranda) formaram o grupo I, enquanto foram as alunas de melhor desempenho durante as primeiras fases de treinamento e as mais engajadas dentro do projeto; C (doutoranda), R (mestre e depois aluna de doutorado) e J (aluna de IC) formaram o grupo II, por serem, entre os restantes, aquelas com o melhor desempenho e o maior engajamento; E, L e P (alunos de IC) e Lu (aluna de mestrado), formaram o grupo III, por serem aqueles que haviam integrado o projeto mais recentemente.

O mesmo texto foi dado aos dez transcritores para que eles o segmentassem, e calculou-se o acordo quanto à anotação das quebras prosódicas entre os integrantes de cada grupo através do teste Kappa (FLEISS, 1971), ou seja, mediu-se o grau de acordo a três nos primeiros dois grupos e a quatro para o terceiro. Depois dos primeiros testes, o grupo I confirmou as expectativas de melhor desempenho; o grupo II obteve um desempenho somente um pouco melhor do que o grupo III. Com base nesses primeiros resultados, resolveu-se reformular o grupo II, excluindo C e R, e incluindo E, L e P do grupo III. Essa decisão foi tomada considerando que E, L e P haviam conseguido resultados muito bons pelo pouco treinamento recebido, e que J parecia ser aquela de melhor desempenho no grupo. Os grupos passaram a ser somente dois, um formado por três transcritores (B, H e M) e outro por quatro (E, J, L e P).

4.2. Os testes de acordo

Para simplificar o cálculo do coeficiente Kappa, as possibilidades de quebras prosódicas consideradas foram três: nenhuma quebra (codificada como "0"), quebra não terminal, equivalente à barra simples e à barra entre colchetes (codificada como "/"), e a quebra terminal, ou seja, a barra dupla ou o símbolo

"+" (codificada como "#").[9] Para o propósito dos testes na fase de treinamento *feedback*, toda fronteira de palavra foi considerada uma posição possível para a anotação de quebra.[10] O quadro 1 a seguir apresenta as possibilidades lógicas de acordo para o grupo 1. Para o Grupo 2, formado por quatro transcritores, obviamente as possibilidades lógicas aumentam, mas seguem o mesmo raciocínio.

QUADRO 1

Possibilidades lógicas de acordo para o Grupo 1

1	acordo total	ausência de quebra (000)
2	acordo total	presença de quebra não terminal (///)
3	acordo total	presença de quebra terminal (###)
4	acordo BH	ausência de quebras contra quebra não terminal de M (00/)
5	acordo HM	ausência de quebra contra quebra não terminal de B (/00)
6	acordo BM	ausência de quebra contra quebra não terminal de H (0/0)
7	acordo BH	ausência de quebra contra quebra terminal de M (00#)
8	acordo HM	ausência de quebra contra quebra terminal de B (#00)
9	acordo BM	ausência de quebra contra quebra terminal de H (0#0)
10	acordo BH	ausência não terminal contra ausência de quebra de M (//0)
11	acordo HM	quebra não terminal contra ausência de quebra de B (0//)
12	acordo BM	quebra não terminal contra ausência de quebra de H (/0/)
13	acordo BH	quebra terminal contra quebra não terminal de M (##/)
14	acordo HM	quebra terminal contra quebra não terminal de B (/##)
15	acordo BM	quebra terminal contra quebra não terminal de H (/##)
16	acordo BH	quebra não terminal contra quebra terminal de M (//#)
17	acordo HM	quebra não terminal contra quebra terminal de B (#//)
18	acordo BM	quebra não terminal contra quebra terminal de H (/#/)
19	acordo BH	quebra terminal contra ausência de quebra de M (##0)
20	acordo HM	quebra terminal contra ausência de quebra de B (0##)
21	acordo BM	quebra terminal contra ausência de quebra de H (0##)
22	desacordo total	(#/0)
23	desacordo total	(#0/)
24	desacordo total	(/0#)
25	desacordo total	(/#0)
26	desacordo total	(0/#)
27	desacordo total	(0#/)

O objetivo era que cada grupo atingisse um escore Kappa de acordo geral a três (ou a quatro) igual ou maior do que 0,8 (o que representaria uma uniformidade excelente na segmentação do *corpus*). Ao mesmo tempo, também esperava-se que o grupo obtivesse um escore Kappa para as quebras terminais mais alto, já que esse tipo de quebra é o mais saliente e mais importante. Adicionalmente, era muito importante que não ocorressem casos dos tipos descritos de 16 a 27 no quadro 1, ou seja, o desacordo extremo em que alguém não percebe nenhuma quebra e alguém percebe quebra terminal. As tabelas 1 e 2 mostram os resultados obtidos pelos grupos em cada texto, na sequência em que foram realizados os testes.[11]

TABELA 1

Acordo da fase de treinamento do grupo 1

	texto	**geral**	**terminal**	**não terminal**
Fase 1	mn01g1	0,75	0,74	0,61
	dl01g1	0,83	0,9	0,62
Fase 2	mn02g1	0,84	0,75	0,77
	dl02g1	0,8	0,85	0,62
Fase 3	mn03g1	0,76	0,71	0,66
	dl03g1	0,78	0,87	0,58

Quanto ao grupo 1, observa-se que o valor de Kappa não aumenta ao longo do processo de testes e discussões. A conclusão é que o grupo 1 já estava suficientemente treinado e que as oscilações são ligadas à dificuldade de cada texto. De fato, em princípio os monólogos devem ser considerados mais difíceis de serem segmentados do que os diálogos. Isso se deve à menor acionalidade e maior textualidade, e consequente aumento no tamanho dos enunciados e da quantidade de quebras prosódicas não terminais.

Além disso, se olharmos para as características de cada texto, podemos observar o seguinte:

a) Os monólogos mn01g1 e mn03g1 são trechos diferentes de uma mesma narrativa, com taxa de elocução extremamente alta, muitas coarticulações e fala muito ritmada. O monólogo mn02g1 possui características opostas.
b) O diálogo dl01g1 é uma interação com alta acionalidade (diálogo entre uma cliente e um vendedor em uma loja de sapatos).

c) O diálogo dl02g1 é uma interação entre dois alunos universitários em que um ensina ao outro o uso de equipamento de gravação. Trata-se de dois falantes de origem rural, com taxa de elocução muito alta e fala muito ritmada.
d) O diálogo dl03g1 consiste em um bate-papo entre duas falantes com alta taxa de elocução e fala ritmada. Dos três diálogos, é o que mais se aproxima das características do monológico, por causa da alternância de turnos longos e baixa acionalidade.

Dadas essas características, não surpreende que o melhor resultado entre os diálogos tenha sido alcançado no dl01g1, e o pior, no dl03g1; e nos monólogos, que o melhor resultado tenha sido conseguido no mn02g1. Vale a pena ressaltar que no mn03g1 foi alcançado um resultado melhor que aquele obtido no mn01g1, mesmo que ambos os monólogos sejam dois trechos diferentes da mesma narrativa.

TABELA 2
Acordo percentual quanto a quebras terminais no grupo 1

segmentação	freq	%
3 quebras terminais	492	66,4%
2 terminais vs 1 não terminal	117	15,8%
1 terminal vs 2 não terminais	111	15,0%
2 terminais vs 1 ausência de quebra	11	1,5%
1 terminal vs 2 ausências de quebra	10	1,3%
Total	741	100,0%

O desacordo extremo (em que pelo menos um segmentador marca terminal e pelo menos um marca ausência de quebra) ocorre em 21 casos no total dos textos do grupo 1. Esse número corresponde a 2,8% do total de ocorrências em que pelo menos um anotador marcou quebra terminal. Analisando esses casos, observamos que são sempre devidos a:

1) Problemas de transcrição. Falta de palavra na transcrição que induz dois segmentadores diferentes a assinalar a quebra em locais diferentes, mas com a mesma intenção.
2) Aspectos de natureza não perceptual, que os segmentadores ainda não tinham clareza sobre como resolver. Um exemplo é a utilização de "hhh", que representa tosse, riso ou outro comportamento paralinguístico. Esses casos não deveriam ser segmentados com qualquer tipo de quebra, mas, na fase de formação, isso ainda não era claro para todos.
3) Erros devidos à distração, imediatamente reconhecidos pelo responsável.

Após as três fases de testes e treinamento, mesmo não sendo alcançado o escore geral de 0,8 em todos os testes, resolvemos que os transcritores estavam prontos para começar o trabalho. Isso porque percebemos que uma quantidade considerável dos desacordos era devida a problemas de transcrição ou à distração imediatamente reconhecida pelo transcritor. Assim, consideramos que o acordo real era maior do que aquele aparente nos testes e, portanto, sempre superior a 0,8. Além disso, deve-se levar em consideração o fato de que cada transcrição do *corpus* (e, portanto, cada segmentação) seria revisada pelo menos duas vezes antes de ser considerada pronta.

Contudo, antes de ser dado o início ao processo de transcrição, foi realizado um teste com o Grupo 1, no qual os segmentadores deveriam assinalar apenas as quebras percebidas como terminais. As quebras prosódicas com valor terminal representam a fronteira mais importante na segmentação da fala, pois delimitam os enunciados. Neste teste foi alcançado um escore kappa de 0,91. Esse resultado representa, praticamente, um acordo de 100% entre os diferentes segmentadores.

TABELA 3
Acordo da fase de treinamento do grupo 2

	texto	**geral**	**terminal**	**não terminal**
Fase 1	dl01g2	0,78	0,81	0,58
	mn01g2	0,73	0,73	0,63
Fase 2	dl02g2	0,76	0,82	0,57
Fase 3	mn02g2	0,68*	0,78*	0,51*
Fase 4	dl03g2	0,78	0,8	0,68
Fase 5	dl04g2	0,77	0,85	0,66
Fase 6	mn03g2	0,77	0,82	0,66
Fase 7	mn04	0,79	0,76	0,7
Fase 8	mn05	0,82	0,83	0,75

* valores calculados com apenas 3 avaliadores

O Grupo 2 realizou oito testes, compreendendo quatro diálogos e cinco monólogos. O primeiro teste consistiu na segmentação de dois textos (dl01g2 e mn01g2), que também foram segmentados pelo grupo 1 (dl01g1 e mn02g1). Os demais testes consistiram em segmentar ou um diálogo, ou um monólogo. A partir de um certo ponto, decidiu-se realizar os testes e o treinamento somente em monólogos, por ser a tipologia mais difícil.

Contrariamente ao que aconteceu com o Grupo 1, no Grupo 2 é evidente a melhora progressiva nos resultados. A queda observada no terceiro teste é aparente, justificada pelo fato de ser o primeiro teste feito com base apenas em um texto monológico, além do fato de ser um teste do qual não participou o segmentador L. Se considerarmos o acordo a três (sem L) também nos outros testes, o acordo total é na maioria das vezes um pouco mais baixo (dl1g2 = 0,77; dl02g2 = 0,76; dl03g2 = 0,79; dl04g2 = 0,76; mn01g2 = 0,74; mn02g2 = 0,68; mn03g2 = 0,75; mn04g2 = 0,78; mn05g2 = 0,81). Se considerarmos o escore Kappa apenas com os segmentadores com maior acordo entre si (L, E e P), os resultados obtidos a partir do terceiro teste são quase sempre melhores (dl1g2 = 0,76; dl02g2 = 0,77; dl03g2 = 0,81; dl04g2 = 0,79; mn01g2 = 0,71; mn02g2 = 0,83; mn03g2 = 0,80; mn04g2 = 0,80; mn05g2 = 0,85).[12] No último teste, o acordo a três sem J chega a 0,85. Isso levou à decisão de que J não realizaria revisões das transcrições que entrariam no *corpus*.

É importante notar que a melhora geral do Grupo 2 se deve a uma melhora tendencial em relação à percepção de quebras terminais, mas principalmente a um aumento constante no acordo em relação às quebras prosódicas não-terminais, que são mais fracas perceptualmente.

TABELA 4
Acordo percentual quanto a quebras terminais no Grupo 2

segmentação	freq	%
3 quebras terminais	400	63,8%
2 terminais vs 1 não terminal	76	12,1%
1 terminal vs 2 não terminais	101	16,1%
2 terminais vs 1 ausência de quebra	35	5,6%
1 terminal vs 2 ausências de quebra	15	2,4%
Total	627	100,0%

Considerando-se os últimos testes, nos quais o grupo 2 obteve os resultados mais consistentes, a porcentagem de desacordo do tipo terminal vs ausência de quebra é de 2,6%. Este valor é ainda um pouco menor do que aquele obtido pelo grupo 1. As mesmas razões já mencionadas para o grupo 1 sobre esse tipo de desacordo também são válidas para o grupo 2.

4.3. Reavaliação do acordo entre transcritores

Depois de um ano de transcrições e revisões, foi realizado um novo teste de acordo entre os transcritores do grupo 1. Esses transcritores trabalharam na transcrição, segmentação e revisão de 30 textos. O propósito desse novo teste era ter uma medida mais precisa do nível de acordo que se encontra realmente na segmentação do (mini)*corpus*.

Com esse teste excluímos um dos fatores de ruído: a má qualidade da transcrição. Isso porque foram utilizados para esse novo teste dois textos transcritos e revisados por um dos membros do grupo 2. O teste reflete o estágio alcançado pelos transcritores no momento da revisão do (mini)*corpus*, fase concluída imediatamente antes de ser realizado o teste.

Os dois textos utilizados para a realização do teste são um diálogo de 562 palavras e um monólogo de 758 palavras. O diálogo consistia em um bate-papo entre duas colegas, uma interação pouco acional, o que apresenta, em princípio, maior dificuldade de segmentação. O monólogo consistia em uma narrativa bastante emocional. Os resultados são os seguintes:

TABELA 5
Acordo nas fases de treinamento e reavaliação do grupo 1

	Treinamento			**Reavaliação**		
tipo de texto	geral	terminal	não terminal	geral	terminal	não terminal
diálogos	0,81	0,87	0,61	0,85	0,86	0,78
monólogos	0,79	0,78	0,69	0,86	0,87	0,78
todos os textos	0,79	0,84	0,66	0,86	0,86	0,78

O escore Kappa geral passa de um intervalo entre 0,75 e 0,84 (geral = 0,79) nos testes da primeira fase para um kappa de 0,86, considerando-se todos os textos juntos. O aspecto especialmente interessante é relativo às medidas desagregadas. Em relação às quebras terminais, passou-se de um intervalo de 0,71 a 0,9 (geral = 0,84) para um valor total de 0,87. O maior progresso se registra em relação à marcação de quebras não terminais. De um intervalo de 0,58 a 0,71 (geral = 0,66) passou-se para 0,78 de acordo geral.

Observamos melhoras evidentes em todos os dados, mas é especialmente evidente a melhora no acordo quanto às quebras fracas (não terminais). Esse resultado pode ser mais bem apreciado analisando duas outras formas de medir

o acordo. A primeira é unir os dois tipos de quebra para avaliar a percepção de quebra prosódica em relação à ausência de quebra, independentemente do tipo de quebra prosódica, terminal ou não terminal. Nos primeiros testes, o acordo geral entre quebra *versus* ausência de quebra foi de 0,86, passando para 0,91 no último teste.

A segunda forma de avaliação diz respeito ao acordo que podemos chamar de realístico, e que considera como base somente as posições em que pelo menos um dos transcritores marcou uma quebra prosódica. Dessa maneira, exclui-se uma grande quantidade de posições (fronteiras de palavras), que não foram marcadas por nenhum dos anotadores, e que apresentam poucos problemas (ou nenhum) no acordo.

TABELA 6
Acordo realístico do treinamento e reavaliação no grupo 1

	Treinamento – k realístico			**Reavaliação – k realístico**		
tipo de texto	geral	terminal	não terminal	geral	terminal	não terminal
diálogos	0,56	0,79	0,46	0,66	0,8	0,65
monólogos	0,51	0,7	0,45	0,63	0,8	0,59
todos os textos	0,55	0,76	0,47	0,65	0,81	0,62

Nos primeiros testes, o resultado para acordo realístico foi 0,55 no acordo geral, 0,76 de acordo terminal e 0,47 para acordo não terminal. Esses dados nos mostram que, mesmo quando o acordo geral não é bom, o acordo sobre as terminais permanece muito bom. Isso mostra a forte saliência perceptual das quebras que definem as unidades básicas da fala, os enunciados.

No último teste temos um kappa realístico geral de 0,66, com 0,81 para as quebras terminais e 0,62 para as quebras não terminais. Se compararmos os dois resultados nessa medida realística, a melhora é muito evidente, e se alcança um escore de acordo que vai de bom para excelente em todos os valores.

5. Considerações finais

A nossa avaliação depois dos testes iniciais foi que os valores de acordo alcançados nos garantiam já um patamar de excelência quanto à confiabilidade estatística da segmentação ao final da fase de treinamento e que teríamos a verificação disso ao final do processo de transcrição com um novo teste, baseado

em uma transcrição confiável e que refletisse o estágio alcançado pelos revisores (que é o que realmente importa para a confiabilidade da segmentação).

O último teste, realizado um ano depois, confirma plenamente as nossas expectativas. A partir desse novo teste, podemos observar com mais detalhe quais são os pontos em que o desacordo foi superado e quais são os pontos em que o desacordo persiste. Em outras palavras, quais são os contextos em que a avaliação perceptual varia mais entre os transcritores. Com base em uma avaliação inicial dos contextos de desacordo, formulamos aqui algumas hipóteses, que serão averiguadas em estudo em andamento.

Eliminando-se os casos de distração e de desacordo relacionados a problemas de transcrição, que obviamente são muito maiores nos primeiros testes, chamaram a nossa atenção três fatores prosódicos extremamente presentes no português do Brasil:

1. A alta frequência de coarticulação e a compatibilidade de quebra prosódica com a coarticulação. Em línguas fortemente silábicas, como o italiano, a quebra prosódica não é compatível com a coarticulação. Ao contrário, no português do Brasil (especialmente na variedade mineira), a compatibilidade entre coarticulação e quebra pode gerar mais dúvidas na percepção.
2. A alta frequência de ênfases, principalmente em unidades tonais muito longas. A ênfase, que é um fenômeno de saliência prosódica, pode facilmente ser confundida com a quebra.
3. A estrutura rítmica do PB. Especialmente em falantes rurais, o padrão rítmico fortemente acentual pode gerar uma impressão semelhante a uma fala escansionada, ou seja, os grupos acentuais podem ser confundidos com unidades tonais.

A avaliação qualitativa dos resultados dos testes, tanto iniciais quanto final, corroborará ou não essas hipóteses.

Antes da publicação do *corpus*, cerca de 20% dos enunciados, escolhidos randomicamente entre todos os textos, serão submetidos à avaliação de não linguistas para um novo teste Kappa sobre o acordo. Esse teste será aplicado e controlado por pesquisadores externos ao projeto.

Notas

[1] O projeto é financiado pela Fapemig, pelo CNPq, pela UFMG e pelo Banco Santander.

[2] A teoria é descrita brevemente na seção 3. Para um resumo da teoria publicado em português, veja-se Raso *et al*. (2007) e Ulisses (2008).

[3] Uma descrição detalhada do *corpus* C-ORAL-BRASIL é apresentada em Raso e Mello, 2009; 2010.

[4] *Retracting* diz respeito a fenômenos de disfluência em que o falante cancela um elemento já pronunciado e o substitui por outro. Ex: Eu vou pro [/1] pra casa de João.

[5] O destaque em negrito foi utilizado para representar as unidades que possuem um foco portador da força ilocucionária, ou seja, as unidades de comentário.

[6] O Kappa, tanto em sua versão tradicional proposta por Cohen (1960) como na modificada por Fleiss (1971) é uma medida de concordância entre observadores. O teste Kappa mede a quantidade de avaliações iguais entre dois (Kappa de Cohen) ou mais (Kappa de Fleiss) juizes, bem como o grau de concordância além do que seria esperado pelo acaso. O valor Kappa pode estar entre 0 e 1, em que 0 significa que o acordo não existe ou é o esperado pelo acaso, e 1 indica um acordo total entre os juízes. Se o Kappa obtido for qualquer valor >0, isto indica que existe alguma concordância. A interpretação dos resultados varia de acordo com o critério adotado pelo pesquisador, contudo encontramos na literatura algumas indicações úteis para a interpretação dos valores do Kappa (LANDIS; KOCH, 1977): 0-0,19 = acordo fraco; 0,2-0,39 = acordo vago; 0,4-0,59 = acordo moderado; 0,6-0,7 acordo substancial; >0,8 = acordo quase total.

[7] Uma descrição detalhada dos critérios de transcrição adotados no C-ORAL-BRASIL é apresentada em Mello e Raso (2009).

[8] O curso de pós de 60 horas e os minicursos de 8 horas foram ministrados por Tommaso Raso; os cursos de pós de 15 horas foram ministrados por E. Cresti e M. Moneglia, assim como os *workshops*. Nem todos os transcritores participaram de todas as atividades de formação. No Grupo 1, todos participaram de todas as etapas, mas a formação do Grupo 2 foi mais heterogênea.

[9] Nos testes, a notação das quebras terminais foi modificada daquela normalmente empregada na transcrição porque, por razões computacionais, havia necessidade de utilizar um símbolo com apenas um caractere para a realização das contagens. Assim, a barra dupla e o símbolo "+" foram ambos convertidos para o símbolo "#".

[10] Em princípio, até fronteira de sílaba pode ser sujeita a quebra em casos excepcionais, como quando silabamos, por exemplo, em meu nome é Leo / nar / do // não

Leopoldo //. Para uma avaliação *a posteriori* do grau de acordo entre os transcritores foram consideradas ainda outras possibilidades, discutidas adiante.

[11] Os textos estão identificados com as siglas dl para diálogos e mn para monólogos, seguidos de número que representam a sequência em que foram segmentados, seguidos da identificação do grupo, "g1" para grupo 1 e "g2" para grupo 2. Os textos segmentados por grupo não foram necessariamente os mesmos.

[12] É importante ressaltar que J foi o único transcritor que fazia parte da formação original do grupo 2. Aparentemente, a partir de certo momento, J também foi superada em aptidão pelos novos membros do grupo, que tinham no início pouquíssima formação.

Referências

AUSTIN, L. J. *How to do things with words*. Oxford: Oxford University Press, 1962. 168 p.

CRESTI, E. Speech act units and informational units. In: FAVA, E. (Ed.). Speech Acts and Linguistic Research. 1994. Buffalo. *Proceedings of the workshop*. Padova: Nemo, 1995. p. 89-107. Disponível em: <http://lablita.dit.unifi.it/preprint/preprint-95coll02.pdf>. Acesso em: 6 set. 2007.

CRESTI, E. *Corpus di Italiano parlato*. Firenze: Accademia della Crusca, 2000. Vol 1, p. 41-166.

CRESTI, E. Per una nuova classificazione dell´illocuzione. In: Convegno SILFI: Tradizione e innovazione, 6. 2000. Duisburg. *Atti*. Pisa: Cesati, 2005. Disponível em: <http://lablita.dit.unifi.it/preprint/preprint-00bcoll01.pdf>. Acesso em: 15 nov. 2008.

CRESTI, E. The informational patterning theory and the corpus based description of spoken language. In: International Workshop in Corpus Linguistics: Bootstrapping Information from Corpora in a Cross Linguistic Perspective, 3. 2008. Firenze. *Proceedings*. Disponível em: <http://lablita.dit.unifi.it/events/cresti.pdf>. Acesso em: 15 nov. 2008.

CRESTI, E.; MONEGLIA, M. (Ed.). *C-Oral-Rom*: Integrated Reference Corpora For Spoken Romance Languages. Amsterdam: John Benjamins, 2005. 303 p.

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

FIRENZUOLI, V. *Le Forme Intonative di Valore Illocutivo dell'Italiano Parlato*: Analisi Sperimentale di un Corpus di Parlato Spontaneo (LABLITA). 2003. 420 f. Tesi (Dottorato in Linguistica) – Università degli Studi di Firenze, Firenze.

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

LANDIS, J. R.; KOCH. G. G. The measurement of observer agreement for categorical data. *Biometrics*, v. 33, p. 159-174, 1977.

MARTIN, P. *WinPitch Corpus*: A text to Speech Alignment Tool for Multimodal Corpora. Lisbon: LREC. May 2004. Disponível em: <http://lablita.dit.unifi.it/coralrom/papers/index.html>. Acesso em: 6 set. 2007.

MELLO, H.; RASO, T. Para a transcrição da fala espontânea: o caso do C-ORAL-BRASIL. *Revista Portuguesa de Humanidades*, v. 13, p. 301-325, 2009.

MONEGLIA, Massimo. The C-ORAL-ROM resource. In: CRESTI, Emanuela; MONEGLIA, Massimo (Org.). *C-ORAL-ROM*: Integrated Reference Corpora for Spoken Romance Languages. Amsterdam: John Benjamis. 2005. p. 1-69.

RASO, T.; MELLO, H. The C-ORAL-BRASIL corpus. In: MONEGLIA, M.; PANUNZI, A. (Org.). *Bootstrapping Information from Corpora in a Cross Linguistic Perspective*. Firenze: Firenze University Press, 2010. p. 193-213.

RASO, T.; MELLO, H. Parâmetros de compilação de um corpus oral: o caso do C-ORAL-BRASIL. *Veredas*, v. 13, p. 20-35, 2009. Disponível em: <http://www.ufjf.br/revistaveredas/files/2009/11/ARTIGO-Tommaso-Raso-e-Heliana-Mello.pdf>. Acesso em: 17 mar. 2010.

RASO, T.; MELLO, H.; DEUS, L.; JESUS, A. Uma aplicação da Teoria da Língua em Ato ao PB. *Revista de Estudos da Linguagem*, v. 15, p. 147-166, 2007.

ULISSES, A. J. *A unidade de Apêndice no português do Brasil*. 2008. 242 f. Dissertação (Mestrado em Linguística) – Faculdade de Letras, Universidade Federal de Minas Gerais, Belo Horizonte.