Filol. Linguíst. Port., São Paulo, v. 20, n. Esp., p. 139-157, 2018
https://doi.org/10.11606/issn.2176-9419.v20iEspecialp139-157

# Para a compilação do C-ORAL-ANGOLA: um corpus de fala espontânea informal do português angolano

## *Toward the compilation of C-ORAL-ANGOLA: an informal spontaneous speech corpus of Angolan Portuguese*

Bruno Rocha[*]
*Universidade Federal do Pará*, Altamira, PA, Brasil

Heliana Mello[**]
*Universidade Federal de Minas Gerais*, Belo Horizonte, MG, Brasil

Tommaso Raso[***]
*Universidade Federal de Minas Gerais*, Belo Horizonte, MG, Brasil

**Resumo:** O trabalho apresenta a arquitetura e os critérios de compilação de um corpus de fala espontânea do português angolano. Após uma breve contextualização da realidade linguística de Angola, são apresentados em detalhe as modalidades de gravação e o tratamento das diferentes variações sociolinguísticas documentadas, destacando-se a atenção à variação diafásica. Em seguida, são detalhados os primeiros 27 textos gravados, que formarão um minicorpus de pelo menos 30.000 palavras, segmentado prosodicamente e oferecendo o texto alinhado ao sinal sonoro. A última parte do artigo é dedicada à discussão dos passos metodológicos da compilação do corpus: definição da qualidade acústica, critérios de transcrição, procedimento de segmentação prosódica, revisão, alinhamento e validação estatística.

**Palavras-chave:** Português angolano. Fala espontânea. Corpus. Compilação.

**Abstract:** The paper introduces the architecture and compilation criteria for an Angolan Portuguese spontaneous speech corpus. After a brief introduction about the linguistic scenario in Angola, we present an in-depth description of the recording modalities and treatment related to the multiple sociolinguistic variations documented, with special attention to diaphasic variation. The first twenty-seven recorded texts are then detailed. These will make up a minicorpus, portraying at least 30,000 words. The minicorpus will be prosodically segmented and will display text-to-speech alignment. The last part of the article is dedicated to the methodological steps taken for the corpus compilation: acoustic quality definition, transcription criteria, prosodic segmentation procedures, revision, alignment and statistic validation.

**Keywords:** Angolan Portuguese. Spontaneous speech. Corpus. Compilation.

FLP20(esp)

[*] Professor Adjunto, Faculdade de Letras, Universidade Federal do Pará, Altamira, PA, Brasil; bbruno791@gmail.com

[**] Professora Titular, Faculdade de Letras, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brasil; hmello@ufmg.br

[***] Professor Titular, Faculdade de Letras, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brasil; tommaso.raso@gmail.com

## 1 INTRODUÇÃO

Neste artigo apresentamos os primeiros textos relativos à compilação de um corpus de fala espontânea do português angolano, pensado segundo os moldes da família C-ORAL, mais precisamente o C-ORAL-ROM (Cresti, Moneglia 2005), para as quatro principais línguas românicas europeias, e o C-ORAL-BRASIL (Raso e Mello 2012 e no prelo) para o português brasileiro (PB).

Entre os dias 10 e 20 de julho de 2018 foram realizadas 28 gravações do português falado em Angola. Entre elas foram escolhidos os textos destinados a entrar no corpus e principalmente aqueles que serão utilizados para compor um minicorpus de português angolano etiquetado informacionalmente com base na *Language into Act Theory* (L-AcT; Cresti 2000; Moneglia, Raso 2014). Chamamos de minicorpus o conjunto de textos destinados a serem etiquetados informacionalmente e a serem inseridos em um corpus mais amplo. Ao longo do artigo, nos referimos portanto ao minicorpus, objeto específico deste texto, e ao corpus como duas entidades distintas, mesmo se fortemente correlacionadas.

L-AcT é uma extensão da teoria dos atos de fala de Austin (1962) que individualiza no enunciado, pragmaticamente e prosodicamente marcado, a interface entre ato locutivo e ilocutivo. A ilocução é a única unidade informacional necessária e suficiente para a a realização do enunciado, mas frequentemente (cerca de 50% dos casos) os enunciados são compostos pela ilocução e outras unidades informacionais não ilocucionárias. As unidades informacionais são tendencialmente isomórficas com as unidades entoacionais. Portanto, um corpus estudável segundo os pressupostos da L-AcT (mas não somente com base nela) precisa possuir pelo menos duas características (aprofundadas ao longo do trabalho): uma forte variação diafásica (o que estimula a emergência da variabilidade ilocucioária e informacional) e uma segmentação prosódica, que marca as fronteiras das unidades entonacionais/informacionais e dos enunciados.

O minicorpus será formado por pelo menos 30.000 palavras, distribuídas em no mínimo 20 textos, e será perfeitamente comparável com os minicorpora já constituídos no Laboratório de Estudos Empíricos e Experimentais da Linguagem (LEEL) da UFMG e no laboratório LABLITA da Universidade de Florença. No LEEL foram desenvolvidos minicorpora de PB informal (Mittmann, Raso 2011), de PB em contexto telefônico (Raso et al. em preparação), e de inglês americano informal (Cavalcante, Ramos 2016); no LABLITA foram desenvolvidos minicorpora de italiano (Panunzi, Mittmann 2014) e de espanhol (Nicolas Martinez, Lombán no prelo). Com a exceção do minicorpus de inglês, os minicorpora são acessíveis através da plataforma IPIC (Panunzi, Gregori 2011), que permite também diferentes tipos de busca nos corpora que a integram.

Ao longo do artigo, para cada aspecto metodológico relativo à compilação do minicorpus, discutiremos em que medida os diferentes recursos da família C-ORAL podem ser considerados comparáveis.

O corpus angolano integra o projeto Libolo[1], coordenado por Carlos Figueiredo da Universidade de Macau e por Márcia Oliveira Santos da USP que, entre outros méritos, tornaram possível essa missão em Angola. Além do apoio e logística proporcionados pelos coordenadores do projeto Libolo, foi muito importante a participação de Graciette Matta, que proporcionou a viabilização de muitas das oportunidades de gravação e deu assistência constante à nossa equipe.

As gravações foram todas realizadas no município do Libolo situado na região do Kwanza Sul, não distante da região da capital Luanda. A maior parte dos textos foi coletada na cidade de Calulo, enquanto outros foram coletados na comuna do Quissongo, uma comunidade rural próxima de Calulo e na comuna de Kabuta. Os falantes gravados são todos falantes de português língua materna ou falantes bilíngues equilibrados de português/kimbundu ou português/kibala, sem que seja possível identificar uma única língua de competência nativa.

## 2 O CONTEXTO LINGUÍSTICO

Segundo os dados do Ethnologue (Simons, Fenning 2018), em Angola estão presentes falantes nativos de 4 grandes famílias linguísticas (dados de 2016): (i) a família indo-europeia, com cerca de 12.300.000 falantes cuja língua nativa é representada quase exclusivamente pelo português; (ii) a família níger-congo com quase 14.000.000 de falantes nativos é representada por 41 línguas; (iii) a família kx'a com pouco mais de 11.000 falantes é representada por 2 línguas; (iv) a família khoe-kwadi com apenas 200 falantes é representada por uma única língua. As duas últimas famílias são limitadas a enclaves presentes apenas no extremo sul do país, mais ou menos próximos à fronteira com a Namíbia.

A região do Libolo, situada ao sudeste da capital, é uma região onde, além do português, se fala o kimbundu ou uma variedade dele chamada kibala ou ngoya (ou identificada através de outros nomes também), própria da transição entre kimbundu e umbundu. Trata-se de uma região prevalentemente cristã, com cidades pequenas e uma ampla área rural. O umbundu é a principal língua africana falada em Angola com cerca de 6.000.000 de falantes nativos (dados de 2012). O kimbundu é língua materna de cerca de 1.500.000 de falantes (dados de 2015). Sua importância é devida também ao fato de ser a língua tradicionalmente falada em Luanda, apesar de a guerra civil ter mudado profundamente a identidade linguística da capital. O kibala, próprio do Libolo, tem, segundo dados de 2000, apenas 2.600 falantes nativos, mas é a variedade banta com a qual se identificavam diversos dos falantes bilíngues gravados, que a chamavam de ngoya.

[1] O projeto *Município do Libolo, Kwanza Sul, Angola: aspectos linguístico-educacionais, históricoculturais, antropológicos e sócio-identitários*, também conhecido como *Projeto Libolo*, é parcialmente financiado pela Universidade de Macau e por entidades privadas filantrópicas de Angola. Trata-se de um projeto internacional e multidisciplinar cujos pesquisadores intervêm, de forma articulada, em pesquisas nas áreas de Linguística, História, Antropologia, Filologia e Acções Pedagógicas. O *Projeto Libolo* está devidanente patenteado pelo Centro de Investigação e Desenvolvimento (R&DAO) da Universidade de Macau, sob o número de referência SRG011-FSH13-CGF, encontrando-se, desta forma, ao abrigo da vigente protecção de direitos autorais de propriedade intelectual designada por "Copyright © 2016, R&DAO University of Macau".

Segundo dados da Central Intelligence Agency relativos a 2015, Angola possui um índice de alfabetização da população a partir dos 15 anos de 71,1% (82% entre os homens e 60,7% entre as mulheres), ocupando o 130º lugar entre 162 países. Como dados comparativos, citamos apenas o Brasil (86º lugar) com 92,6%, Portugal (62º lugar) com 95,7% e a média mundial com 86,2%.

## 3 MODALIDADES DE GRAVAÇÃO

As gravações foram realizadas em duas modalidades técnicas distintas, dependendo do número dos participantes de cada sessão. Na primeira modalidade, com apenas um ou dois falantes principais, foram utilizados um gravador Marantz (pmd 660) ou Tascam (DR-100 MKII) e microfones de lapela sem fio (Transmitters Bodypack Transmitter SK 100 G3 e Receivers Diversity Receiver EK 100 G3), permitindo assim que os falantes se locomovessem com liberdade durante o período de gravação, resultando portanto em gravações de uma maior variedade de situações. Na segunda modalidade, quando os falantes principais eram mais de dois, era utilizado também um *mixer* (Behringer Xenyx 1222fx) para além dos equipamentos já descritos, a fim de permitir o uso de mais de dois microfones para os dois canais de gravação. O número máximo de microfones usados foi seis, mas em algumas gravações se superou esse número de participantes. Nesse caso os microfones foram posicionados de modo a aumentar as probabilidades de gravar todas as vozes com a melhor qualidade possível. Todo o equipamento era móvel, com a exceção do *mixer*. Isso significa que apenas as gravações de conversações (diálogos com mais de dois participantes principais) obrigavam os falantes a estarem a uma distância de não mais de 30-50 metros do gravador, que não podia ser movido por estar ligado ao *mixer*. Nas gravações com 2 microfones o gravador podia ser movido no caso de os participantes se afastarem do ponto de início da gravação.

Por *falantes principais* entende-se aqueles falantes que estavam previstos na fase de planejamento da gravação e aos quais foram aplicados os microfones. Contudo, nas situações de fala espontânea em um contexto natural é frequente que durante a situação planejada para a gravação se insiram outros falantes não previstos. Quando isso acontece, os falantes não previstos podem ou não ser captados adequadamente pelos microfones. Isso, somado a outros fatores, condiciona a qualidade da gravação. Em situações específicas, aconteceu de os falantes se afastarem do gravador em direções opostas por alguns minutos. Nesses casos o gravador não podia ser movido, já que a direção do movimento dos falantes gravados era distinta; as consequências, dependentes das decisões tomadas pelos pesquisadores no momento em que isso ocorria foram várias; as diversas decisões tomadas, finalizadas a minimizar os danos à qualidade acústica, foram as seguintes: excluir momentaneamente um ou mais falantes da gravação, desligando um ou mais microfones (essa foi a decisão tomada tipicamente em casos de conversações com muitos falantes); isolar um canal para evitar que o afastamento de um dos falantes gerasse ruído que comprometesse a fala do outro falante principal, que continuava interagindo com falantes sem microfones (essa decisão foi frequente em situação como aquelas em que dois vendedores de uma loja ou de um mercado, portadores dos microfones, interagiam com clientes, e um dos falantes principais se afastava, ou em situações comparáveis). Em geral, as gravações duraram muito tempo (em média entre uma e duas horas), tornando

possível a recuperação de um trecho suficientemente longo com qualidade acústica adequada.

Os textos que serão transcritos para composição do minicorpus terão uma duração média de 1.500 palavras, ou seja, pouco mais ou pouco menos de 10 minutos, dependendo da tipologia textual, do grau de interação, das quantidades de silêncio propiciadas pela situação e da velocidade de fala dos falantes. Em nenhum caso os textos do minicorpus serão significativamente maiores que esse marco, para evitar a falta de balanceamento; em alguns casos os textos poderão ser menores, mas sempre salvando a integridade textual.

Vale uma observação de ordem ética. Na realidade angolana, e ainda mais em uma cultura substancialmente tradicional e rural como aquela do Libolo, não é possível apresentar aos falantes o termo de consentimento que é elaborado por um comitê de ética a ser lido e assinado, como aconteceu no caso de todos os corpora da família C-ORAL. Quem concede a permissão para a gravação e transmite aos falantes a garantia de confiabilidade nos pesquisadores, em geral, é o *Soba*. O *Soba*, assistido pelos *Sobetos*, é de fato a maior autoridade civil da comunidade, desde os tempos pré-coloniais. Ele exerce a função de ligação entre a comunidade e o governo. Mesmo em centros maiores, cada bairro possui o seu *Soba*. O *Soba Grande* tem autoridade sobre os *Sobas* de uma determinada região. De fato, antes de começar as gravações, nos encontramos com o *Soba Grande* de Calulo, para nos apresentar e pedir a autorização para as gravações. A autorização nos foi concedida e o *Soba Grande* tornou-se, inclusive, um dos falantes em uma de nossas gravações. As gravações futuras, que serão realizadas em grandes centros urbanos, serão acompanhadas por um termo de consentimento nos moldes dos outros corpora.

FLP20(esp)

## 4 A VARIAÇÃO DIAFÁSICA

O corpus tem como um de seus objetivos principais retratar a variação diafásica da fala angolana. A primeira divisão interna do corpus é em três grandes modalidades comunicativas: monólogos, diálogos e conversações (esta com mais de 2 e um máximo de 8 falantes principais, alcançando apenas o número máximo de 6 no minicorpus). Cada modalidade será representada no minicorpus com pelo menos 10.000 palavras. Para o corpus o objetivo é de cerca de 50.000 palavras por modalidade, de modo a alcançar um corpus de pelo menos 150.000 palavras. Se essa é a proporção ideal para representar as três modalidades (e constitui o objetivo do trabalho), o que importa mais é manter uma proporção de um terço de fala monológica e dois terços de fala dialógica (diálogos e conversações), já que a diferença estrutural entre conversações e diálogos é pequena (Raso, Mittmann 2012; Cresti 2005)

Dentro da modalidade monológica se buscou variação entre gêneros textuais: explicações profissionais, relatos de experiências de vida, relatos de eventos ligados à história recente ou à cultura do lugar. Dentro das modalidades dialógica e conversacional se buscou a maior variedade situacional em função da maior variedade acional. Ao variarem a modalidade de fala e a situação, variam também a tipologia de atos de fala eliciados e a estruturação informacional dos enunciados, permitindo assim que sejam coletados dados com uma variação não limitada apenas ao nível morfossintático e lexical. Portanto, os pesquisadores buscaram gravar a

maior variação possível de situações comunicativas, evitando a repetição da mesma situação e situações pouco acionais e repetitivas como bate-papos e entrevistas. As pessoas gravadas nas modalidades dialógica e conversacional estavam sempre empenhadas em uma atividade específica, como mostraremos mais à frente. É isso que garante o alto grau de interatividade e acionalidade das gravações da família C-ORAL, explicitamente desenhada para o estudo das ilocuções e da estruturação informacional em contexto natural (Moneglia 2005; Raso 2012; Raso, Mello 2014).

## 5 AS OUTRAS VARIAÇÕES

A variação *diatópica* do minicorpus já foi indicada previamente. Contudo, se esse minicorpus reflete apenas a fala do Libolo, o corpus maior é destinado a refletir a fala de uma região mais ampla e com prevalência clara (pelo menos 50%) da fala de Luanda, assim como os outros corpora da família C-ORAL escolheram a diatopia de uma grande área urbana (Madri, Marselha, Florença, Lisboa, Belo Horizonte).

Quanto à variação diastrática, o minicorpus (na medida do possível) e o corpus (com maior rigor) buscam equilibrar em número de palavras a fala masculina e aquela feminina, assim como os falantes das diversas faixas de escolarização e das diferentes faixas etárias. A distribuição das faixas de escolarização e de idade é ainda objeto de discussão, pois Angola não possui uma distribuição comparável àquela dos outros países da família C-ORAL[2]. Será portanto necessário conciliar as exigências de comparabilidade com os outros corpora com aquelas de representatividade da sociedade angolana. Em princípio, está sendo seguido o critério adotado no C-ORAL-BRASIL: três faixas de escolarização (1: até o primeiro grau incompleto; 2: até o terceiro grau, mas não usado na ocupação exercida; 3: superior) e cinco faixas etárias (M: menor de idade; A: até 25 anos; B: até 40 anos; C: até 60 anos; D: mais de 60 anos). A indicação do sexo é marcada com F (feminino) e M (masculino). Quando um dado é desconhecido, é marcado com 'X'.

FLP20(esp)

Nos metadados aparece também a ocupação profissional e a origem específica dos falantes, além da descrição da situação, do lugar e do tópico da interação.

Tomou-se especial cuidado para que houvesse diversidade de falantes no minicorpus. Nenhum falante poderá ultrapassar 1.700 palavras. Se um falante aparece em mais de uma gravação, o que acontece em apenas dois casos, será considerada a soma das palavras, nunca superior a 1.700. Nas gravações algumas poucas vezes aparece também a fala dos pesquisadores (que são brasileiros com a exceção de um italiano). As palavras dos pesquisadores não serão levadas em conta na contagem mínima de palavras.

[2] O próprio C-ORAL-BRASIL modificou um pouco a indicação das faixas de escolaridade, pelo fato de o Brasil apresentar um quadro um pouco diferente daquele da realidade europeia.

## 6 Os textos

### 6.1 Monólogos

1. *Experiência de guerra*. O falante conta para os pesquisadores a própria experiência na guerra civil como responsável da artilharia no sul do país. Falante de Luanda: sexo M; idade C; escolaridade 2; ocupação: dono de um escritório de despachante. Qualidade acústica B (veja a seção 7 sobre a qualidade acústica.
2. *Passeio*. Um jovem do lugar acompanha um dos pesquisadores em uma rápida visita no centro de Calulo. Falante 1 de Calulo: sexo M; idade B; escolaridade 2; ocupação: auxiliar lingüístico do projeto. Falante 2 brasileiro: sexo M; idade B; escolaridade 3; ocupação: professor universitário. Qualidade acústica AB.
3. *Monólogo no hotel*: A falante explica aspetos da própria profissão aos pesquisadores. Falante 1 de Calulo: sexo F; idade D; escolaridade 2; ocupação: oficial de notário. Falante 2 brasileira: sexo F; faixa etária C; escolaridade 3; ocupação: professora universitária. Falante 3 italiano: sexo M; faixa etária C; escolaridade 3; ocupação: professor universitário. Qualidade acústica AB.
4. *Saudade do Libolo*. Um falante da antiga elite colonial conta a própria parábola de vida ao pesquisador. Falante 1 de Calulo (que deixou aos 17 anos para Portugal e para onde voltou 20 anos depois): sexo M; faixa etária C; escolaridade 3; ocupação: contábil. Falante 2 italiano: sexo M; faixa etária C; escolaridade 3; ocupação: professor universitário. Qualidade acústica AB.
5. *Na escola*. Dois pequenos monólogos de professores de escola em reunião com visitantes do projeto Libolo. Falante 1 de Gabela (Ambuim, Kwanza Sul): sexo F; idade B; escolaridade 2; ocupação: professora de ensino primário. Falante 2 de Kabuta (Libolo): sexo F; idade B; escolaridade 2; ocupação: professora de ensino pré-escolar. Falante 3 de Calulo: sexo F; idade B; escolaridade X; ocupação: professora. Falante 4 de Calulo: sexo: M; idade C; escolaridade 2; ocupação: professor de ensino primário. Falante 5 brasileiro: idade C; escolaridade 3; ocupação: professor universitário. Qualidade acústica B.
6. *Passeio na fazenda*. (Kabuta) O falante conta uma história de guerra acontecida no lugar. Falante de Calulo: sexo M; idade B; escolaridade 3; ocupação: funcionário do município. Qualidade acústica: A.
7. *No Kissongo*. Falante de Calulo: sexo F; idade A; escolaridade 2; ocupação: supervisora de compras na fazenda. Qualidade acústica C.

### 6.2 Diálogos

1. *Atendimento médico 1*. Um médico atende alguns pacientes. Falante 1 de Lubango (Huila); sexo M; idade B; escolaridade 3; ocupação: médico. Falante 2 de Calulo: sexo F; idade M; escolaridade 1. Qualidade acústica AB.
2. *Atendimento médico 2*. O médico é o mesmo do *Atendimento médico 1*. Falante 1 de Lubango (Huila); sexo M; idade B; escolaridade 3; ocupação: médico. Falante 2 de Calulo: sexo F; faixa etária M; escolaridade 1. Qualidade acústica AB.
3. *Pasteleiros*. Dois confeiteiros que trabalham em um restaurante preparando pães e bolos a serem servidos em um evento no dia seguinte. Falante 1 de

FLP20(esp)

Calulo; idade: B; escolaridade: 1; ocupação: confeiteiro. Falante 2 de Calulo; idade: B; escolaridade: 2; ocupação: confeiteiro. Qualidade acústica AB.

4. *Balneários*. Duas faxineiras limpam os vestiários do estádio de Calulo depois de um jogo de futebol. Falante 1 de Calulo: idade B; escolaridade 1; ocupação: faxineira do clube Libolo. Falante 2 de Kabuta (Libolo): idade B; escolaridade 1; ocupação: faxineira do clube Libolo. Qualidade acústica BC.
5. *Lavando o carro*. Dois jovens de Calulo lavam carros a pagamento no Rio de Calulo. Falante 1 do Quissongo (Libolo): sexo M; idade A; escolaridade 2; ocupação: lavador de carro. Falante 2 do Quissongo (Libolo): sexo M; idade A; escolaridade 1; ocupação: lavador de carro. Qualidade acústica B. Durante a gravação outros lavadores de carro intervêm rapidamente. Qualidade acústica AB.
6. *Cadastro no hospital*. Dois atendentes do hospital conversam entre si enquanto fazem cadastro dos pacientes. Falante 1 de Mussafo (Malanje); sexo: M; idade B; escolaridade 2; ocupação: atendente no hospital de Calulo. Falante 2 de Calulo; sexo: F; faixa etária B; escolaridade 2; ocupação: atendente no hospital de Calulo. Qualidade acústica AB. Alguns pacientes aparecem rapidamente na interação.
7. *Mercado*. Uma cozinheira de restaurante vai ao mercado e negocia com uma vendedora. Falante 1de Calulo; idade B; escolaridade 1; ocupação: cozinheira de restaurante. Falante 2 de Calulo; idade B; escolaridade 1; ocupação: vendedora no mercado de Calulo. Qualidade acústica AB.

### 6.3 Conversações

1. *Dominó*. Um grupo de jovens joga um jogo de dados típico da região, que é chamado de *dominó*, na frente da casa de um deles. Falante 1 de Calulo: sexo M; idade A; escolaridade 1; ocupação: estudante. Falante 2 de Calulo: sexo M; idade M; escolaridade 1; ocupação: estudante. Falante 3 de Calulo: sexo M; idade A; escolaridade 1; ocupação: estudante. Falante 4: sexo M; idade M; escolaridade 1; ocupação: estudante. Qualidade acústica B.
2. *Lanche*. 5 Faxineiras lanchando em pausa do serviço. Falante 1 de Calulo: sexo F; idade B; escolaridade 2; ocupação: faxineira. Falante 2 de Calulo: sexo F; idade A; escolaridade 2; ocupação: faxineira. Falante 3 de Luanda: sexo F; faixa etária B; escolaridade X; ocupação: faxineira. Falante 4 de Calulo: sexo F; idade B; escolaridade 2. Falante 5 de Calulo: idade C; escolaridade 2; ocupação: faxineira. Qualidade acústica C.
3. *Discoteca*. Três jovens de Calulo conversam enquanto desmontam o equipamento de uma discoteca. Falante 1 de Calulo: sexo F; idade B; escolaridade 2; ocupação: secretária. Falante 2 de Calulo; sexo M; idade B; escolaridade 2; ocupação: dono de discoteca. Falante 3 de Calulo; sexo M; idade B; escolaridade 2; ocupação: colaborador linguístico do projeto. Qualidade acústica B.
4. *Montando os gols*. Três funcionários do Clube Recreativo Desportivo do Libolo montam pequenas traves para treinos de futebol. Falante 1 de Calulo; sexo M; idade A; escolaridade 1; ocupação: funcionário de serviços gerais. Falante 2 de Calulo; sexo M; idade B; escolaridade 1. Falante 3 da Uíge; idade C; escolaridade 1; ocupação: funcionário de serviços gerais. Qualidade acústica B.
5. *Funcionários da fazenda Cleonas*. Três funcionários da fazenda Cleonas conversam após o fim do expediente. Falante 1 de Calulo; sexo M; idade C;

FLP20(esp)

escolaridade 2; ocupação: administrador da fazenda. Falante 2 de Bangu-Uanga; idade D; escolaridade 1; ocupação: funcionário da fazenda. Falante 3 de Calulo; sexo M; idade D; escolaridade 1; ocupação: funcionário da fazenda.

6. *Conversa na escola*. Gravação realizada na escola da Missão Católica de Calulo, em um encontro com um professor brasileiro para discutir questões sobre a juventude em Calulo. Falante 1 de Calulo; sexo F; idade A; escolaridade 1; ocupação: estudante. Falante 2 de Calulo; sexo M; idade A; escolaridade 1; ocupação: estudante. Falante 3 do Kwanza Norte; sexo F; idade M; escolaridade 1; ocupação: estudante. Falante 4 de Dondo (mas mudou para Calulo no primeiro ano de vida); sexo M; idade A; escolaridade 1; ocupação: estudante. Falante 5 de Calulo; sexo M; idade A; escolaridade 1; ocupação: estudante. Falante 6 de Calulo; sexo F; idade A; escolaridade 1; ocupação: estudante. Falante 7 de Calulo; sexo M; idade A; 1; ocupação: estudante. Falante 8 de Calulo; sexo M; idade A; escolaridade 1; ocupação: estudante. Falante 9 brasileira; sexo F; idade A; escolaridade 2; ocupação: estudante. Falante 10 brasileiro; sexo M; idade C; escolaridade 3; ocupação: professor universitário. Qualidade acústica B.
7. *Cozinha da pousada*. Falante 1 de Calulo: sexo F; idade C; escolaridade 1; ocupação: chefe de cozinha. Falante 2 de Calulo: sexo F; idade C; escolaridade 1; ocupação: camareira. Falante 3 de Calulo: sexo M; idade A; escolaridade 2; ocupação: garçom. Falante 4 de Calulo: sexo M; idade A; escolaridade 2. Qualidade acústica AB.
8. *Embalando presentes*. Quatro amigas embalam presentes. Falante 1 de Calulo; sexo F; idade C; escolaridade 3; ocupação: conselheira e gestora hoteleira. Falante 2 de Luanda; sexo F; idade C; escolaridade 1; funcionária de serviços gerais no hotel. Falante 3 de Calulo; sexo F; idade C; escolaridade 1; ocupação: camareira de hotel. Falante 4 brasileira; idade C; escolaridade 3; ocupação: professora universitária. Qualidade acústica BC.
9. *Papelaria*. Os dois donos de uma papelaria interagem com os clientes.Falante 1 de Calulo; sexo M; idade B; escolaridade X; ocupação: dono de papelaria. Falante 2 de Calulo; sexo M, idade B; escolaridade X; ocupação: dono de papelaria. Qualidade acústica B.
10. *Cozinha na fazenda*. Conversa na cozinha do restaurante da Kabuta. Falante 1 da Kabuta; sexo M; idade B; ocupação: recepcionista. Falante 2 de Mucula dos Dambos; sexo F; idade B; escolaridade 1; ocupação cozinheira. Falante 3 do Libolo; sexo M; idade B; escolaridade 1; ocupação: cozinheiro. Falante 4 de XXX; idade X; escolaridade X; ocupação: dono do restaurante. Qualidade acústica B.
10. *Soba*. O *Soba Grande* de Calulo recebe a equipe do projeto e, junto com tia Ká, explica como é eleito o Soba e como são a vida e a morte de um Soba. Dois professores estrangeiros intervêm para fazer perguntas e comentários. Falante 1 de Calulo: M; faixa etária D; escolaridade 1; ocupação: Soba. Falante 2 de Calulo; F; faixa etária C; escolaridade 3; ocupação: conselheira e gestora hoteleira. Falante 3 de Calulo (mas vive fora de Calulo desde os 19 anos de idade); M; faixa etária D; escolaridade 3; ocupação: professor universitário. Falante 4 brasileira; F; faixa etária C; escolaridade 3; ocupação: professora universitária. Qualidade acústica AB.
11. *Volta no mercado*. Uma cozinheira e uma vendedora, que são amigas, dão uma volta nas várias lojas do mercado. Falante 1 de Calulo; faixa etária B; escolaridade 1; ocupação: cozinheira de restaurante. Falante 2 de Calulo; faixa etária B; escolaridade 1; ocupação: vendedora no mercado de Calulo.

FLP20(esp)

Outros falantes: donos de outras lojas. Não estão disponíveis os dados sobre estes falantes. Qualidade acústica B.

12. *Cozinhando na fazenda*. Conversação durante a preparação do almoço na fazenda Cleonas. Falante 1 de Kabuta (Libolo): sexo M; idade B; escolaridade 2; ocupação: recepcionista do restaurante. Falante 2 Mucula dos Dambos (Libolo); idade B; escolaridade 1; ocupação: cozinheira. Falante 3 do Libolo; idade B; escolaridade 1; ocupação: cozinheiro. Falante 4 de X; idade X; escolaridade X; ocupação: proprietário do restaurante. Qualidade acústica B.
13. *Regando a grama*. Três funcionários do Clube Recreativo Libolo regam a grama depois do jogo. Falante 1 de Calulo; sexo M; idade A; escolaridade 1; ocupação: funcionários de serviços gerais. Falante 2 de Calulo; sexo M; idade A; escolaridade 1; ocupação: funcionários de serviços gerais. Falante 3 de Calulo; sexo M; idade B; escolaridade 1; ocupação: funcionários de serviços gerais. Qualidade acústica C.

## 7 A QUALIDADE ACÚSTICA

A qualidade acústica foi classificada nas opções A (melhor qualidade), AB, B, BC, C (pior qualidade aceita para o corpus). A avaliação leva em conta os seguintes critérios, com base na classificação de Raso (2012), integrada parcialmente com os critérios de Carrenho, Constantini, Barbosa (2017), considerando que este último trabalho classifica os áudios para finalidades diferentes das nossas:

a) verificação da possibilidade de escuta;
b) cálculo da relação sinal ruído do áudio, que deve ser feito em, no mínimo, dois pontos do arquivo (em trecho com maior presença de ruído e em trecho com menor concentração de ruído, escolhidos pela observação da forma de onda);
b) verificação da possibilidade de cálculo da curva de frequência fundamental ($f_0$);
c) verificação da possibilidade de cálculo dos dois primeiros formantes nas vogais;
d) identificação das fricativas e de sua concentração de energia;
e) verificação da presença de ruído de fundo;
f) verificação da presença de trechos com sobreposição de voz.

Uma qualidade ideal deve permitir a análise dos formantes e uma curva de $f_0$ confiável para quase toda a gravação. Uma gravação com o mínimo de aceitabilidade deve permitir a extração da $f_0$ confiável para pelo menos 60% da gravação e ter uma boa resposta dos microfones. A tolerância é menor para os monólogos, média para os diálogos (que dependem fortemente da situação de gravação) e mais alta para as conversações (que, além de depender das características da situação, inevitavelmente levam a uma quantidade maior de sobreposições). É importante considerar que a variação situacional é objetivo prioritário do corpus, e portanto é inevitável aceitar gravações com qualidade acústica não ideal.

Apresentamos a seguir imagens em *Praat* (Boersma, Weenink 2018) através de duas figuras. A Figura 1 mostra um enunciado típico na qualidade muito alta (A) e alta (AB), e a Figura 2 mostra um enunciado típico nas outras qualidades. As imagens

ilustram principalmente a relação do sinal de voz com o ruído de fundo; os outros critérios levam à classificação específica dentro dos dois grandes grupos.

Mais especificamente, um áudio avaliado como de qualidade muito alta ou alta possui quase sempre uma qualidade apropriada para quase todo tipo de análise fonética, poucas sobreposições de voz, quase nenhum ruído de fundo, computação da $f_0$ possível em (quase) todo o arquivo, calculabilidade dos dois primeiros formantes das vogais, boa ou média identificação das fricativas e da concentração de energia das mesmas. A relação sinal-ruído do áudio é acima de 20 dB e frequentemente alcança ou supera 30 dB.

As qualidades média e baixa indicam um áudio com uma boa quantidade de trechos apropriados para a análise fonética e, no mínimo, 60% dos trechos com um cálculo confiável da $f_0$. São possíveis algumas dificuldades na identificação das fricativas e, em alguns casos, no cálculo do F2 das vogais. A escuta é sempre clara, com exceções muito localizadas. As sobreposições podem ser frequentes, mas sem comprometer os critérios mínimos mencionados. A relação sinal-ruído de fundo pode variar muito abaixo dos 20 dB, chegando em alguns trechos a ser até inferior a 10 dB.

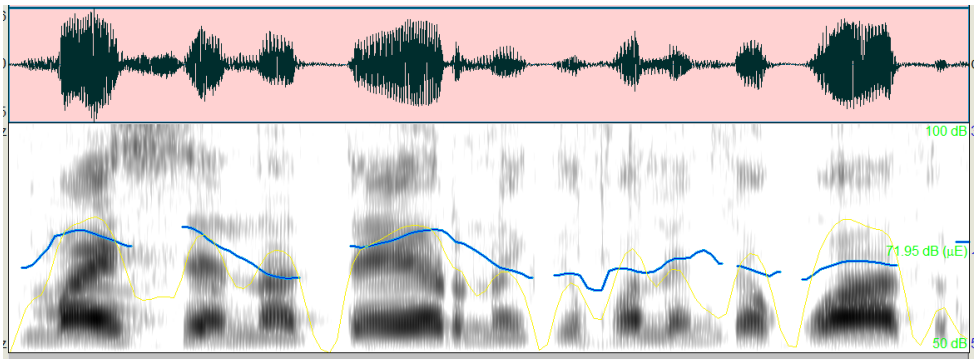


Figura 1 - Espectrograma de um enunciado com qualidade A ou AB

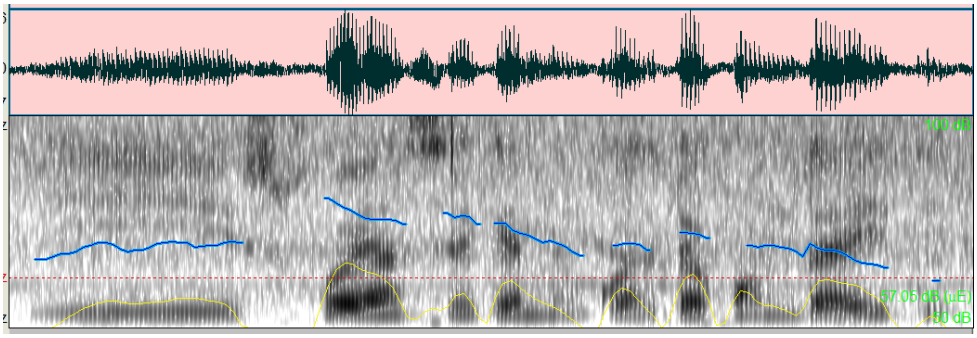


Figura 2 - Espectrograma de um enunciado com qualidade B, BC ou C

Contudo, em gravações como estas, a indicação da qualidade acústica deve ser tomada como uma síntese das características acústicas da gravação, e não necessariamente como uma referência constante. De fato, em contexto natural as condições acústicas são sujeitas a mudanças contínuas: dependendo das atividades, certos ruídos podem ser constantes ou não; em gravações com posição variável dos falantes o contexto acústico pode mudar rapidamente e constantemente; as sobreposições podem ser frequentes ou muito localizadas; as sobreposições podem

se concentrar em alguns trechos ou serem distribuídas ao longo da interação. Portanto, se em alguns casos o julgamento da qualidade acústica pode ser tomado como uma característica mais ou menos constante do texto, em outros casos o mesmo texto apresenta características acústicas muito variáveis, e a qualidade deve ser considerada mais como uma média aproximada de características, até muito diferentes, que coexistem no texto. Essa observação é importante para quem busca somente trechos de qualidade alta: o fato de a gravação ter sido etiquetada como de qualidade média ou baixa não significa que não possua trechos de qualidade alta.

## 8 TRATAMENTO DOS DADOS

Os dados serão tratados como nos casos dos outros corpora da família C-ORAL, tentando melhorar os aspectos qualitativos, como sempre tem ocorrido ao longo dos mais de dez anos de atividade do LEEL. As fases do tratamento dos dados são descritas em 8.

### 8.1 A transcrição

Os textos serão transcritos com base na lógica adotada principalmente nos corpora de italiano e de PB (Mello et al. 2012). Nesses corpora os critérios preveem que, a partir da base dos critérios ortográficos, sejam individualizados os fenômenos potencialmente em curso de gramaticalização e lexicalização, e que esses casos sejam transcritos não ortograficamente a fim de que possam ser recuperados automaticamente através de um *parser* (vejam-se Bick 2012 e 2014).

FLP20(esp)

Os critérios devem contudo garantir um equilíbrio entre exigências diversas: a legibilidade dos textos, a recuperabilidade dos fenômenos interessantes, as exigências do *parser*, a necessidade de consistência por parte dos segmentadores. Esta última exigência merece talvez uma explicação, que pode melhor ser oferecida através de um exemplo: o PB apresenta formas dos pronomes pessoais plenas e reduzidas em todas as pessoas; contudo, se é fácil distinguir com coerência as variantes da segunda e da terceira pessoa (tanto no singular quanto no plural) *você(s)/ocê(s)/cê(s)*; *ele(s)/e*, *es*; *ea(s)*, não parece possível manter coerência nas formas ditongada e variamente monotongadas da primeira pessoa singular, ou nas formas com a vogal e sem a vogal da primeira pessoa plural (quando *nós* pode ser pronunciado com variantes aproximadamente constituídas por uma sibilante precedida por uma nasalização); portanto se mantem sempre as formas *eu* e *nós*. Nas primeiras pessoas é muito difícil encontrar acordo entre os transcritores e até entre momentos diferentes do mesmo transcritor. Isso não contribui para a recuperabilidade do fenômeno e apenas gera confusão; portanto é melhor renunciar a preservar o fenômeno através de um critério não ortográfico.

No caso do corpus angolano os critérios adotados para o corpus de PB podem constituir uma boa base de partida, mas quase certamente deverão ser modificados em parte. De fato as duas variedades não compartilham todos os fenômenos que podem ser considerados potencialmente em curso de lexicalização ou gramaticalização; em alguns casos, formas que são candidatas interessantes a serem diferenciadas em PB podem ser transcritas ortograficamente na variedade angolana, a qual, por outro lado, apresentará outros fenômenos que merecem uma transcrição não ortográfica. As decisões a esse respeito podem ser tomadas somente depois de

uma primeira transcrição de diversos textos e com base em uma discussão que envolve todos os transcritores. Durante a fase de transcrição do C-ORAL-BRASIL, com sete transcritores, se chegou a fechar os critérios somente depois de cerca de seis meses. Provavelmente, dada a experiência adquirida, o processo para este corpus poderá ser mais rápido, mas devemos ser prudentes a esse respeito, considerando que a maioria dos transcritores não será falante nativa da variedade angolana do português.

### 8.2 A segmentação prosódica

É amplo o consenso na comunidade científica que um nível importante da organização da fala é constituído pelo agrupamento de poucas (às vezes apenas uma) palavras em unidades chamadas de unidades tonais, unidades entonacionais, grupos prosódicos ou com outras denominações (Barth-Weingarten 2016; Barbosa, Raso 2018; Izre'el et al. no prelo). A essas unidades, dependendo da teoria adotada, é atribuído um preciso valor funcional ou cognitivo. Essas unidades são separadas por fronteiras nitidamente perceptíveis pelos falantes (os testes mostram um acordo claramente superior a 80%, inclusive na fala espontânea). Na abordagem teórica que adotamos (que não é necessária para se utilizar o corpus), a unidade entonacional é tendencialmente isomórfica com a unidade informacional, incluindo a unidade que carrega a função ilocucionária. Portanto a segmentação prosódica é considerada essencial para os nossos estudos. Mas recentemente a segmentação prosódica tem se tornado quase a norma na compilação de corpora de fala espontânea (além dos corpora da família C-ORAL, vejam-se Du Bois et al. 2000-2005, Mettouchi et al. 2015, Izre'el e Rahav 2004, entre outros).

FLP20(esp)

A segmentação prosódica, de fato, não é importante apenas para quem atribui valor funcional linguístico às unidades entonacionais, mas parece o elemento proeminente para identificar unidades necessárias para delimitar um âmbito de análise da sequência de fala, frequentemente chamadas de unidades de referência, do ponto de vista comunicativo na fala em contexto natural. De fato, nós precisamos construir as relações linguísticas para interpretar os enunciados, e segmentar a fala é crucial para isso. A segmentação feita apenas a partir de pausas é completamente inconfiável no caso da fala espontânea, como mostrado amplamente na literatura (Raso et al 2015; Mittmann e Barbosa 2016, entre outros). Por exemplo, uma sequência como *João vai pro Rio até amanhã* pode ser segmentada como um ou mais enunciados:

- João vai pro Rio (asserção ou pergunta, ou outro ato comunicativo) // até amanhã (despedida) //
- João (chamamento ou pedido de confirmação ou outro ato) // vai pro Rio até amanhã (ordem ou pedido de confirmação ou outro) //
- João (chamamento ou outro) // vai pro Rio (ordem ou outro) // até amanhã //

Esses exemplos mostram como é importante segmentar a fala para definir o âmbito em que acontecem as relações linguísticas. A mesma sequência sintático-semântica pode adquirir muitos valores comunicativos diferentes dependendo de informações que são de natureza exclusivamente prosódica e em que as informações de fronteira são decisivas, mesmo se se combinam como informações prosódicas de outra natureza.

Quanto à unidade que deve ser tratada como referência, ou seja como âmbito das principais relações linguísticas, alguns autores privilegiam a unidade entonacional (Mettouchi et al. 2015), outros uma unidade entonacional ou um conjunto delas que se conclua com uma fronteira de um tipo específico, ou seja, uma fronteira que carrega a percepção de conclusão (Izre'el no prelo; Cresti 2000). Para estes últimos, a percepção de fronteira deve ser acompanhada de um valor ilocucionário e da percepção de terminalidade para que se possa constituir uma unidade de referência. Essa é também a nossa proposta.

Portanto, na segmentação, distinguimos entre uma fronteira não terminal (/) e uma fronteira terminal (//). Dois outros símbolos completam a anotação prosódica: o símbolo (+) indica enunciado interrompido (seja por motivo interno ou externo ao falante) e o símbolo ([/n]) indica retratação (o número associado à barra entre colchetes indica o número de palavras retratadas).

### 8.3 O alinhamento do texto ao som

Os corpora de fala de terceira geração apresentam todos o alinhamento do texto ao som. Se na primeira geração se considerava suficiente trabalhar nas transcrições, e se na segunda geração o áudio acompanha as transcrições porém sem nenhum alinhamento, agora se considera essencial que o som seja alinhado ao texto para que a fala possa ser realmente estudada (Mello, 2014). De fato, somente nesse caso podemos utilizar as informações veiculadas pelo canal sonoro tantas vezes quanto acharmos necessário e com extrema facilidade. Não podemos esquecer que a fala é um processo, e não um produto como a escrita; a fala, portanto, desaparece imediatamente, e a única maneira para observá-la é repetir através de meios tecnológicos o processo dela. Anexar o áudio a um corpus transcrito, sem o alinhamento, não produz uma diferença significativa para o estudo da fala, que de fato continuará a se basear somente, ou quase somente, na transcrição, ou seja, em um texto que tem sua origem na fala, mas que não é mais fala, mas sim escrita, tendo portanto perdido todas as informações do canal sonoro, *in primis* a prosódia (Linell, 2005).

### 8.4 A revisão

Uma vez transcrito, segmentado e alinhado, o corpus deverá ser revisado. As fases de transcrição e segmentação, por serem ambas de natureza perceptual, podem ser (e normalmente são) realizadas em concomitância. A revisão pode ser realizada depois dessa fase ou depois da fase de alinhamento, segundo o que se achar mais oportuno para o andamento do trabalho. Contudo a fase de revisão de transcrição e segmentação é delicada. Principalmente quanto à revisão da segmentação (mas, em medida menor, também quanto à fase de transcrição), a revisão normalmente não deve ser feita por todos os que participaram das primeiras fases. O corpus deve alcançar o maior grau possível de consistência, e nem todos temos a mesma percepção prosódica, a mesma capacidade de não atribuir à fronteira prosódica fenômenos perceptuais que podem ser devidos a outros objetivos que não são os de marcar fronteira (os casos mais clássicos são a confusão entre proeminência e fronteira e a confusão entre fronteira sintática e fronteira prosódica) e a mesma atenção na aplicação dos critérios de transcrição. É provável portanto que alguns componentes do grupo de pesquisa demonstrem uma maior sensibilidade para uma

tarefa ou para outra. Geralmente é aconselhável se utilizar um teste Kappa de Fleiss (1971) para definir a melhor estratégia para a fase de revisão, identificando os segmentadores com maior consistência. Normalmente uma única revisão não é suficiente e frequentemente são necessárias três ou quatro fases de revisão, para garantir que a fase de validação seja bem sucedida.

**8.5 A validação estatística**

Tanto a fase de segmentação quanto aquela de transcrição devem ser validadas. A validação da segmentação é principalmente uma validação prévia, ou seja, uma validação da capacidade dos segmentadores em realizar a sua tarefa. A validação das transcrições é principalmente uma validação *a posteriori*, ou seja, uma validação dos resultados alcançados.

No C-ORAL-BRASIL, um grupo de potenciais segmentadores (ou mais frequentemente apenas uma parte deles) foi considerado pronto somente depois de ter alcançado um acordo superior a 0,8 em um teste Kappa de Fleiss (1971). Antes das revisões a seleção foi mais rígida: não contava apenas o acordo geral, mas 0,8 era o objetivo mínimo não somente para o acordo geral mas também para o acordo relativo a cada tipo de fronteira (terminal e não terminal) (cf. Mello et al., 2012). Normalmente, o acordo para as terminais é significativamente maior do que aquele para não terminais. Por experiência, os dois acordos tendem a ser parecidos somente quando o resultado é especialmente bom. O acordo geral entre os revisores do C-ORAL-BRASIL foi de 0,86 (0,87 para as terminais e 0,86 para as não terminais), o que é considerado excelente.

Na validação das transcrições do C-ORAL-BRASIL foi necessário considerar em separado: (a) cada critério não ortográfico (que normalmente são muitos e constituem portanto um grupo amplo de validações); (b) os critérios ortográficos em conjunto; (c) a acurácia das marcas de fronteira (por exemplo a presença dos colchetes nas retratações e a real correspondência dos números associados às marcas prosódicas com as palavras realmente canceladas pelo falante); (d) a acurácia em marcar as palavras interrompidas (marcadas pelo símbolo '&' antes da palavra); (d) a quantidade de erros por enunciado (cf. Mello et al., 2012).

O C-ORAL-BRASIL se deu como objetivo que nenhum critério ultrapassasse 5% de erros. O grupo (a) normalmente é o mais desafiador, pois a quantidade de erros não é homogênea para todos os fenômenos, e porque a maior dificuldade para os transcritores reside exatamente na aplicação dos critérios não ortográficos. Se um ou mais dos critérios ultrapassarem esse limiar, o corpus inteiro deve ser novamente revisado, limitadamente aos critérios que apresentaram resultados insatisfatórios. O primeiro C-ORAL-BRASIL, dedicado à fala informal, não precisou de uma nova revisão após a validação, mas o segundo C-ORAL-BRASIL, dedicado à fala formal, a mídia e a telefone, precisou que os transcritores revisassem novamente todas as 300.000 palavras do corpus em busca de erros relativos a um pequeno grupo de fenômenos. Naturalmente, isso acarretou uma segunda fase de validação que garantisse que essa última revisão tivesse resolvido os problemas identificados na primeira validação.

A validação, dado o seu custo de tempo, não pode ser realizada no corpus inteiro, mas uma revisão devida a uma validação insatisfatória deve ser feita sobre o corpus inteiro, mesmo se limitadamente aos fenômenos com erros superior a 5%.

FLP20(esp)

A metodologia seguida no C-ORAL-BRASIL é a seguinte: são extraídos aleatoriamente 10% dos enunciados de cada texto, e são analisados para cada um dos critérios. O que acontece nesses casos é que alguns critérios apresentam um número de ocorrência suficiente para uma avaliação considerada significativa dos erros (um mínimo de 50 ocorrências) e outros não. O primeiro grupo é portanto avaliado, tomando-se desde já uma decisão sobre a necessidade ou não de uma revisão posterior. Para todos os outros casos, se extraem novamente 10% de enunciados de cada texto (naturalmente não coincidentes com a primeira amostra) e se procede da mesma maneira. Pode acontecer de alguns fenômenos ainda não alcançarem uma quantidade de ocorrências significativa. Nesse caso é necessário distinguir entre aqueles que estão próximos desse limite e aqueles que são de frequência tão baixa que se deveria fazer a validação sobre o corpus quase inteiro. No primeiro caso se justifica um aumento tardio da amostra; no segundo provavelmente não.

Mas o mais importante é que o usuário do corpus saiba o grau de confiabilidade dos critérios para cada fenômeno. Se um usuário desejar, por exemplo, fazer uma pesquisa sobre as duas séries (plena e reduzida) de formas pronominais, ele deve saber qual é a margem de erro que o corpus apresenta e poder decidir se para seus objetivos é uma margem aceitável ou não. Contudo, os erros não são todos iguais. Por exemplo, no caso do C-ORAL-BRASIL, nós distinguimos entre as formas *vamos*, *vamo* e *vão*. Analogamente distinguimos entre formas verbais com normalização do sufixo verbal (tal como a alternância *es fazem*/*es faz*). Casos como o primeiro são fáceis de se buscar automaticamente. Qualquer problema na acurácia da transcrição pode facilmente ser resolvido pelo usuário do corpus. Bem diferente é o segundo caso, porque os lexemas aos quais o fenômeno se aplica são muitos e não previsíveis. Essa é uma outra reflexão a se fazer quando se avalia se é o caso ou não de assumir um certo custo em termos de tempo e trabalho humano para aperfeiçoar a transcrição. Um problema que pode facilmente ser enfrentado pelo usuário é naturalmente menos grave que um problema que o usuário não pode corrigir, e que, portanto, deve ser resolvido na fase de compilação do corpus.

FLP20(esp)

### 8.6 A etiquetagem

O minicorpus angolano, assim como todos os minicorpora da coleção C-ORAL, será etiquetado informacionalmente com base na L-AcT (Moneglia, Raso 2014; Cresti 2000). A etiquetagem é um processo manual, realizado por etiquetadores treinados, e produz uma anotação que permite estudos sobre a estruturação informacional comparáveis entre as diversas línguas e entre as diversas tipologias textuais anotadas (para o PB e o italiano veja-se Panunzi, Mittmann 2014. Outros estudos estão em curso sobre o espanhol e o inglês americano). Já existe uma plataforma especializada para a consulta dos minicorpora anotados no laboratório LABLITA (Panunzi, Gregori 2011; http://www.lablita.it/app/dbipic/); em breve o laboratório LEEL permitirá a consulta também em uma plataforma própria.

## 9 CONCLUSÃO

Neste artigo apresentamos pela primeira vez o projeto de um corpus de fala espontânea do português angolano e os textos já gravados e que estão sendo tratados para a realização de um minicorpus de pelo menos 20 textos e 30.000 palavras (mas provavelmente mais) segmentado prosodicamente e etiquetado informacionalmente. Ainda não existe, até onde seja do nosso conhecimento, um corpus de fala espontânea do português angolano. Os dados que se tornarão disponíveis com a realização do corpus aqui anunciado, e já com o minicorpus, representam portanto uma contribuição importante para o estudo científico dessa variedade do português de maneira comparável com o PB, retratado nos corpora C-ORAL-BRASIL, e, mesmo se em medida menor, com o PE, retratado no corpus C-ORAL-ROM ou em outros corpora (Santos e Freitas 2008; Bettencourt Gonçalves e Veloso 2000; Bacelar do Nascimento 2001).

## REFERÊNCIAS

Bacelar do Nascimento F, editora. Português falado - documentos autênticos: gravações áudio com transcrições alinhadas. Lisboa: Centro de Linguística da Universidade de Lisboa e Instituto Camões; 2001. [citado 17 dez. 2018]. Disponível em: http://clul.ulisboa.pt/equipa/fbacelar/portugues_falado_2001_nascimento.pdf

Barbosa PA, Raso T. Spontaneous speech segmentation: functional and prosodic aspects with applications for automatic segmentation. Revista de Estudos da Linguagem. 2018;26(4):1361-1396.

Barth-Weingarten D. Intonation units revisited caesura in talk-in-interaction. Amsterdam: John Benjamins; 2016.

Bettencourt Gonçalves J, Veloso R. Spoken Portuguese: geographic and social varieties. Proceedings of the Second International Conference on Language Resources and Evaluation. Volume II. Athens, Greece: National Technical University of Athens Press; 2000. p. 905-908.

Bick E. A anotação gramatical do C-ORAL-BRASIL. In: Raso T, Mello H, editores. C-ORAL-BRASIL I. Corpus de referência do português brasileiro falado informal. Belo Horizonte: UFMG; 2012. p. 223-254.

Bick E. The grammatical annotation of speech corpora. Techniques and perspectives. In: Raso T, Mello H, editores. Spoken corpora and linguistic studies. Amsterdam: John Benjamins; 2014. p. 105-128.

Boersma P, Weenink D. Praat: doing phonetics by computer [programa de computador]. Amsterdam: Universiteit van Amsterdam; 2018. [citado 17 dez. 2018]. Disponível em: http://www.fon.hum.uva.nl/praat.

Carrenho JM, Constantini AC, Barbosa PA. Qualidade acústica para análises na fonética forense: construção de uma proposta de classificação. Comunicação ao XXIV Congresso Nacional de Criminalística, VII Congresso Internacional de Pericial Criminal, XXIV Exposição de Tecnologias Aplicadas à Criminalística.

Cavalcante F, Ramos A. The American English spontaneous speech minicorpus: architecture and comparability. CHIMERA: Romance Corpora and Linguistic Studies. 2016;3(2):99-124. [citado 17 dez. 2018]. Disponível em: https://revistas.uam.es/index.php/chimera/article/view/6507.

Central Intelligence Agengy. The world factbook. [citado 5 out. 2018]. Disponível em: https://www.cia.gov/library/publications/the-world-factbook/fields/2103.html.

FLP20(esp)

Cresti E. Corpus di italiano parlato. Firenze: AccademiadellaCrusca; 2000. 2 Vols.

Cresti E. Notes on lexical strategy, structural strategies and surface clause indexes in the C-ORAL-ROM spoken corpora. In: Cresti E, Moneglia M, editores. C-ORAL-ROM: integrated reference corpora for spoken Romance Languages. Amsterdam, Philadelphia: John Benjamins; 2005. p. 209-256.

Cresti E, Moneglia M, editores. C-ORAL-ROM: integrated reference corpora for spoken Romance Languages. Amsterdam, Philadelphia: John Benjamins; 2005.

Du Bois J W, Chafe WL, Meyer C, Thompson S, Santa Barbara Corpus of Spoken American English. Washington DC: Linguistic Data Consortium; 2000-2005.

Fleiss JL. Measuring nominal scale agreement among many raters. Psychological Bulletin. 1971;76:378-382.

Gregori L, Panunzi A. DB-IPIC: An XML database for informational patterning analysis. In: Mello H, Pettorino M, Raso T, editors. Proceedings of the 7th GSCP International Conference. Speech and Corpora. Florence: Firenze University Press; 2012. p. 121–127.

Izre'el S. Syntax, prosody, discourse and information Structure: the case for unipartite clauses. A View from Spoken Israeli Hebrew. Revista de Estudos da Linguagem; no prelo.

Izre'el S, Mello H, Panunzi A, Raso T, editores. In search for a reference unit of spoken language: a corpus driven approach. Amsterdam: John Benjamins; em preparação.

Izre'el S, Rahav G. The corpus of spoken Israeli Hebrew (CoSIH); Phase I: the pilot study. In: Oostdijk N, Kristoffersen G, Sampson G, editors. LREC 2004 Sattelite Workshop, Fourth International Conference on Language Resources and Evaluation: Compiling and Processing Spoken Language Corpora. Lisbon, Portugal. Paris: ELRA - European Language Resources Association; 2004. p. 1-7.

Linell P. The written language bias in linguistics. New York: Routledge; 2005.

Mello H. Methodological issues for spontaneous speech corpora compilation. The case of C-ORAL-BRASIL. In: Raso T, Mello H, editores. Spoken corpora and linguistic studies. Amsterdam: John Benjamins; 2014. p. 27-68.

Mello H, Raso T, Mittmann M, Vale H, Côrtes P. Transcrição e segmentação prosódica do corpus c-oral-brasil: critérios de implementação e validação. In: Raso T, Mello H, editores. C-ORAL-BRASIL I. Corpus de referência do português brasileiro falado informal. Belo Horizonte: UFMG; 2012. p. 125-174.

Mettouchi A, Vanhove M, Caubet D, editors. Corpus-based studies of lesser-described languages: the CorpAfroAs corpus of spoken Afro Asiatic languages. Studies in Corpus Linguistics 68. John Benjamins: Amsterdam-Philadelphia; 2015.

Mittmann MM, Barbosa PA. An automatic speech segmentation tool based on multiple acoustic parameters. CHIMERA: Romance Corpora and Linguistic Studies. 2016;3(2):133-147.

Mittmann MM, Raso T. The C-ORAL-BRASIL informationally tagged minicorpus. In: Mello H, Panunzi A, Raso T. Pragmatics and prosody: illocution, modality, attitude, information structure and speech annotation; 2011. p. 151-183

Moneglia M. 2005. The C-ORAL-ROM resource. In: Cresti E, Moneglia M, editors. C-ORAL-ROM: Integrated reference corpora for spoken romance languages. Amsterdam: John Benjamins; 2005. p. 1–70.

FLP20(esp)

Moneglia M, Raso T. Notes Language into Act Theory (L-AcT). In: Raso T, Mello H, editors. In: Spoken Corpora and Linguistic Studies. Amsterdam: John Benjamins; 2014. p. 468-495.

Nicolas Martinez C, Lombán M. Mini-Corpus del español para DB-IPIC. CHIMERA. Romance Corpora and Linguistic Studies. No prelo.

Panunzi A, Gregori L. DB-IPIC. An XML database for the representation of information structure in spoken language. In: Mello H, Panunzi A, Raso T, editors. Pragmatics and prosody. Illocution, modality, attitude, information structure and speech annotation. Florence: Firenze University Press; 2011. P. 19–37.

Panunzi A, Mittmann MM. The IPIC resource and a cross-linguistic analysis of information structure in Italian and Brazilian Portuguese In: Raso T, Mello H, editors. Spoken corpora and linguistic studies. Amsterdam: John Benjamins; 2014. p. 129-151.

Raso T. O corpus C-ORAL-BRASIL. In: Raso T, Mello H, editores. C-ORAL-BRASIL I Corpus de referência do português brasileiro falado informal; 2012. 55–90.

Raso T, Mello H, editores. C-ORAL-BRASIL I. Corpus de referência do português brasileiro falado informal. Belo Horizonte: UFMG; 2012.

Raso T, .Mello H. C-ORAL-BRASIL: description, methodology and theoretical framework. In: Tony Berber Sardinha T, São Bento TL, editors. Working with Portuguese Corpora. London-New Delhi-New York-Sydney: Bloomsbury; 2014. p. 257-278.

Raso T, Mello H, editores. C-ORAL-BRASIL I. Corpus de referência do português brasileiro da fala formal em contexto natural, de mídia e de telefone. Em preparação.

Raso T, Mittmann MM. As principais medidas da fala. In: Raso T, Mello H, editores. C-ORAL-BRASIL I. Corpus de referência do português brasileiro falado informal. Belo Horizonte: UFMG, 2012. p. 177-220.

Raso T, Mittmann MM, Oliveira A. O papel da pausa na segmentação prosódica de corpora de fala. Revista de Estudos da Linguagem, v. 23; 2015. p. 883-922-922. Disponível em: http://www.periodicos.letras.ufmg.br/index.php/relin/article/download/9536/8799.

Raso T, Soares E, Miranda I. Um minicorpus de fala telefônica do português brasileiro etiquetado informacionalmente; em preparação.

Santos F, Freitas T. CORP-ORAL: Spontaneous speech corpus for European Portuguese. In: Proceedings of the International Conference on Language Resources and Evaluation, LREC; 2008. Disponível em: http://www.lrec-conf.org/proceedings/lrec2008.

Simons GF, Fenning CD, editors. Ethnologue: languages of the world, languages of Angola, Twenty-first edition. Dallas, Texas: SIL International; 2018. Disponível em: www.ethnologue.com.

FLP20(esp)