The C-ORAL-BRASIL Project

Coordinators: Tommaso Raso and Heliana Mello

Subprojects in progress:

- Compilation of a BP spontaneous speech corpus
- Study of: main speech measurements, information structure, and BP illocutions, based on the Informational Patterning Theory (CRESTI 2000) (T. Raso)
- Study of modality and its expression in BP (H. Mello)
- Comparative studies in grammaticalization and lexicalization between BP and EP with a view to language contact phenomena (H. Mello and T. Raso)

The C-ORAL-BRASIL CORPUS

This is the fifth leg of the C-ORAL-ROM (CRESTI-MONEGLIA 2005); and as the C-ORAL-ROM it comprises the following:

- at least 30 hours and 300.000 words of spontaneous speech, divided in at least 200 texts, half informal and the other half formal and phone interactions;
- the transcription should be done following the CHILDES CLAN system (implemented through prosodic annotation);
- utterance and tone unit segmentation (MONEGLIA-CRESTI 1997);
- F0-sound-text alignment through the WinPitch software (Ph. MARTIN);
- Information unit tagging (following the Informational Patterning Theory (CRESTI 2000);
- Morphosyntactic and POS tagging and concordance lists through the PALAVRAS computational grammar (E. BICK).

Some advantages of the *corpus*

- immediate comparability, due to its architecture and annotation criteria, to the languages documented in the C-ORAL-ROM (EP, Italian, French and Spanish);
- systematic documentation of diaphasic variation and partial documentation of the diastratical variation of the Mineiro dialect;
- segmentation into utterances and tonal units;
- immediate accessibility to sound to text alignment and to spectrograms;

•information unit tagging database;
•morphosyntactic POS tagging (E. BICK);
• concordances;
• speech measurement calculations for (based on time and number of words): number of utterance; number of tonal units; number of retractings and interruptions (utterances and words); number of utterances with and without verbs; number and placing of negation particles and conjunctions (*e* 'and', *mas* 'but', *porque* 'because', *que* 'that' (beginning of utterance, beginning of unit, within unit, dedicated unit).

Phase 1 (2007-2009): the informal *corpus*

- Objective: at least 100 spontaneous speech texts: 1/3 monologues; 1/3 dialogues; 1/3 conversations (80% in family/private contexts and 20% in public contexts). Recording, transcription, segmentation, alignment, informational and morphosyntactic tagging.
- Equipment: digital recorder Marantz PMD660; kits wireless sennheiser evolution EW100 ENG G2 (receiver, transmitter, clip on microphone, batteries); omnidirecional microphones sennheiser MD421; mixer Xenyx 1622 (6 channels).

What we have right now

•Recordings: around 220 (text number is superior to that), most in stereo. Recordings are still being done despite the fact that we have reached our goal.
• Segmented transcriptions: around 80, revisions around 30, alignment around 15, informational tagging 5.
• Balanced 5,000 utterance mini-corpus for preliminary studies.

Next steps

• Finish up transcriptions, revisions and alignment.
• Go on with tagging.

The diaphasic variation

-It allows us to study speech structural variations, based on the following oppositions:

• formal vs. informal; public vs. family/private; dialogic interaction vs. monologic interaction;

- It also allows the collection of different speech acts, based on the multiple activities performed by informants.

Examples: people grocery shopping and shoe shopping; construction worker and an engineer at a construction site; people playing cards; a student helping another one with a recorder; driver and passenger talking in a car; waiters waiting at a party; 4 people playing soccer and pool; a mother telling a story to her child; people telling dramatic moments of their life or explaining their job; jokes; recipes, etc.

Interlinguistic comparability

It allows us to check:

- shared speech features and particular features of a given language;
- the specificities of a Romance language outside Europe, such as BP, in comparison with European Romance languages;
- features likely to result from contact phenomena, in a comparison between BP and EP.

Transcription criteria

Text sizes: 70% will have around 1,500 words; 10% will have around 4,500 words; 20% will have less than 1,000 (only if textually autonomous).

Transcriptions are done by expert transcribers (usually people who made the recordings); later they are revised by another transcriber; a second revision is made during the alignment process; a third one during the informational tagging.

Statistical validation for the prosodic agreement among 3 transcribers shows a kappa always over 0,80 (over 0,90 for terminal breaks).

Transcriptions are orthographic, with some modifications to represent some specially significant speech aspects such as:

- lack of plural markings: *os menino bonito 'the-PL boy-SG handsome-SG'*;
- plural marking in invariable words: *ques menino bonito 'what-PL boy-SG handsome-SG'*;
- subject cliticization: tonic *você, ele 'you, he'* vs. clitic *cê (cês), e'* (*ea, es, eas*);
- reduction of demonstratives (*aque´'that-MASC'; aquea 'that-FEM', daques 'of those-MASC', etc.)*
- contraction of articulated prepositions: *pro, pra, pros, pras 'for the'*; *co, ca, cos, cas 'with the'*; *dum, duma, duns, dumas 'of the', etc.*
-apheresis: *tá, tava, tando, etc. (< estar 'be')*; *güento (< agüento 'stand'), pera (<espera 'wait'), etc.*

- reduction of the verbal paradigm (*nós faz < nós fazemos "we do"; es diz < eles dizem "they say";* etc.);
- serial verbs (*ele foi falou "he went said"; ele pegou falou "he took said"*, etc.)
-apocope: expressions such as *po' fazer < pode fazer "(you) can do (it)", o' <*olha "look";
- diminutive forms: *sozim <sozinho "alone"*, *certim <certinho "right-DIM"*, *etc.*;
- exclamations: *Nossa < No' "Our Lady", Vixe'* <Virgem Maria "Virgin Mary";
- loss of copula in interrogative and cleft constructions (*que que cê fez < o quê é que você fez "what did you do"; por que que cê veio < por que é que você veio "whay did you come"; ele que veio < ele é que veio "he was the one who came", etc.*)
- cliticization of negation;
- etc.
-The corpus will have a glossary of non-orthographic forms and the criteria followed. In principle, only features that have a chance of being grammaticalized or lexicalized are marked.

Metadata

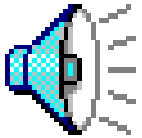
Each text has the following set of metadata:
-Title;
-File name (eg. bfamcvo1 means that it is the first text in the Brazilian-family-conversation group);
-Acoustic qualify (A, AB, B, BC or C) defined by parameters such as instruments used, environment noise level, number of overlappings, percentage of legible or reliable prosodic curve;

- Recording length in time units (only the transcribed and aligned parts);
- Number of words;
- Place of recording;
- Class: eg. Informal: family_private: dialogue;
- Subject (what is being talked about);
- Setting (eg. Friend being interviewed at home during lunch time, microphone not hidden, researcher interacts with subject);
-Participants. For each participant it is necessary to point out:
• first name and three letter abbreviation (eg. BAO, Bruno) in order to indicate turns in the transcription;

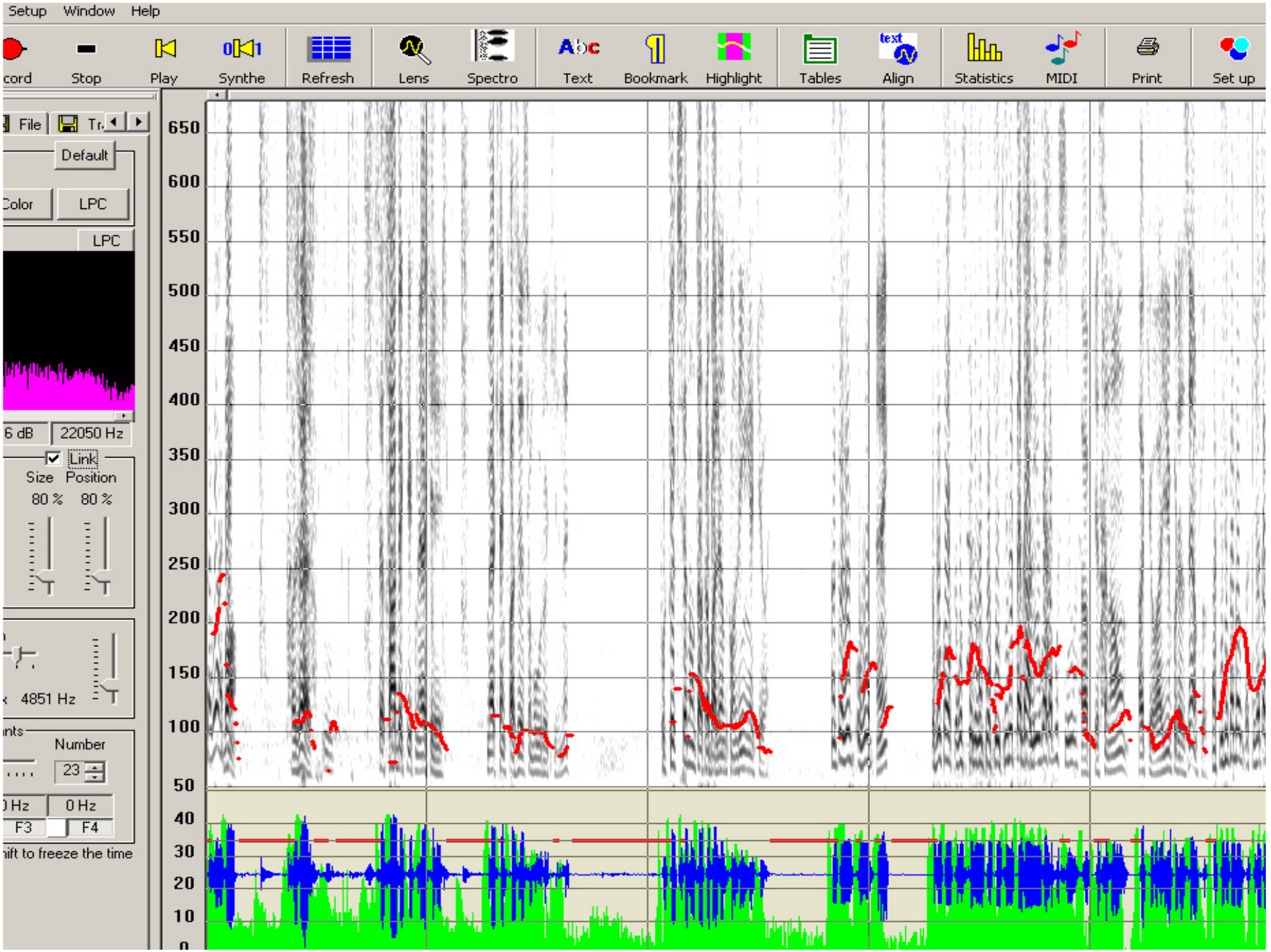
• role (if he/she takes part in the interaction, if he/she is an interviewer, and what participation he/she has);
• origin;
• gender;
• age (A = 18-25 years old, B = 26-40, C = 41-60, D = more than 60, M = underage, X = unknown);
• schooling (1 = no schooling or up to grade school, 2 = high school graduate or college graduate whose profession does not require the title, 3 = college graduate whose profession require the title or post graduate);
• occupation.

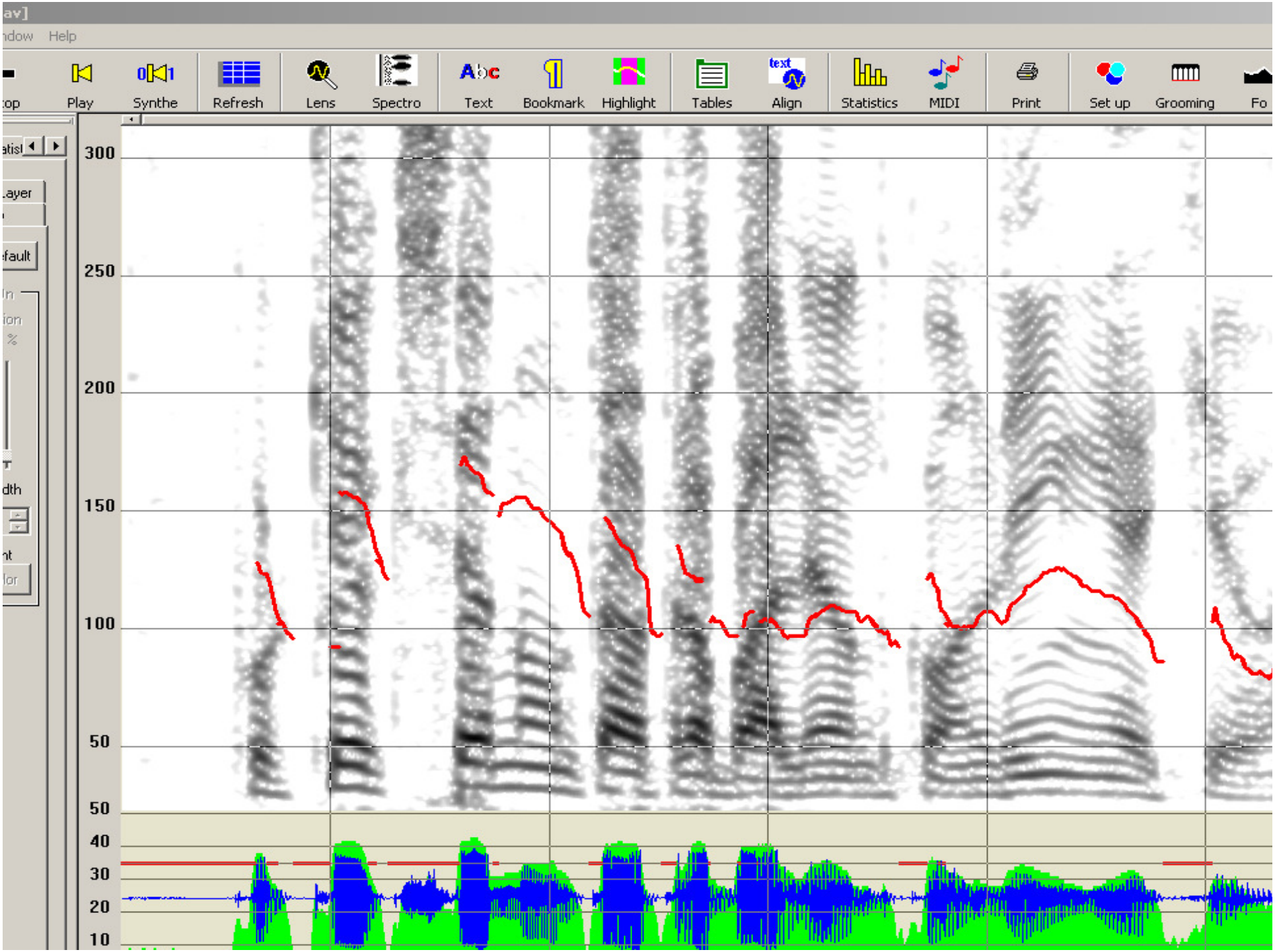
All participants sign an agreement form, approved by the Ethics Committee.

Example: transcription and segmentation



*PAU: bom // Rogério //
*ROG: hum //
*PAU: cê sabe aqui como é que [/3] como é que tem que fazer esse muro aqui / né // por que que cê não tá trabalhando com linha aqui o' //
*ROG: ah / então eu vou [/2] eu vou &f +
*PAU: hein //
*ROG: eu vou &coloc [/3] eu vou suspender mais um pouquinho aqui / e vou pegar a linha e colocar por cima //
*PAU: ah / porque se não + aqui o' // aí / por exemplo +
*ROG: aqui já tá dando [/4] aqui já tá dando a altura //
*PAU: o' aqui + não // tá dando a altura daquele que a Isa /
*ROG: é //
*PAU: / marcou lá / <né> //
*ROG: <que a dona> Isa marcou ali //


Setup
Window
Help
cord
Stop
Play
Synthe
Refresh
Lens
Spectro
Text
Bookmark
Highlight
Tables
Align
Statistics
MIDI
Print
Set up
File
Default
Color
LPC
LPC
6 dB
22050 Hz
Link
Size
Position
80 %
80 %
4851 Hz
Number
23
0 Hz
F3
F4
ift to freeze the time
650
600
550
500
450
400
350
300
250
200
150
100
50
40
30
20
10
0


av]
ndow Help
op
Play
Synthe
Refresh
Lens
Spectro
Text
Bookmark
Highlight
Tables
Align
Statistics
MIDI
Print
Set up
Grooming
Fo
300
250
200
150
100
50
50
40
30
20
10

Segmentation criteria

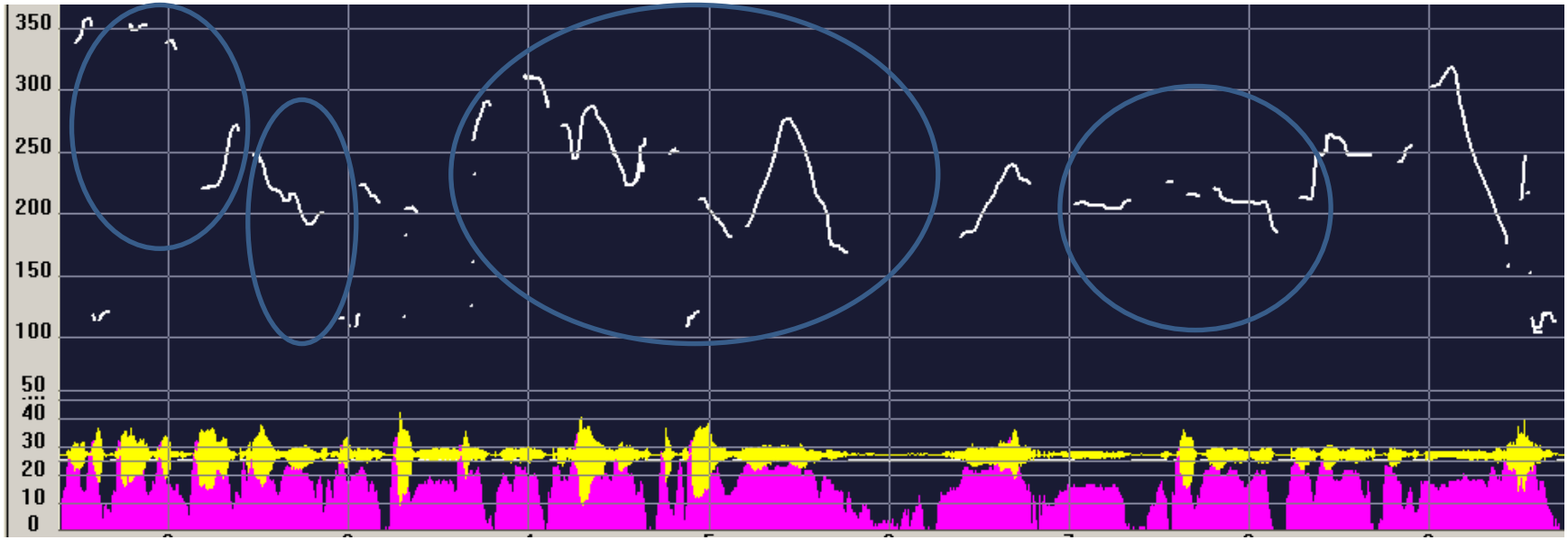
Terminal break (//): indicates the end of an utterance, that is, the smallest autonomous interpretable unit.

Non-terminal break (/):indicates tonal unit within an utterance – in principle, it is an information unit.

Retracting ([/n]): some execution problem – it could have a partial or total repetition, or no repetition. The n indicates the number of words cancelled by the speaker and shouldn’t be counted.

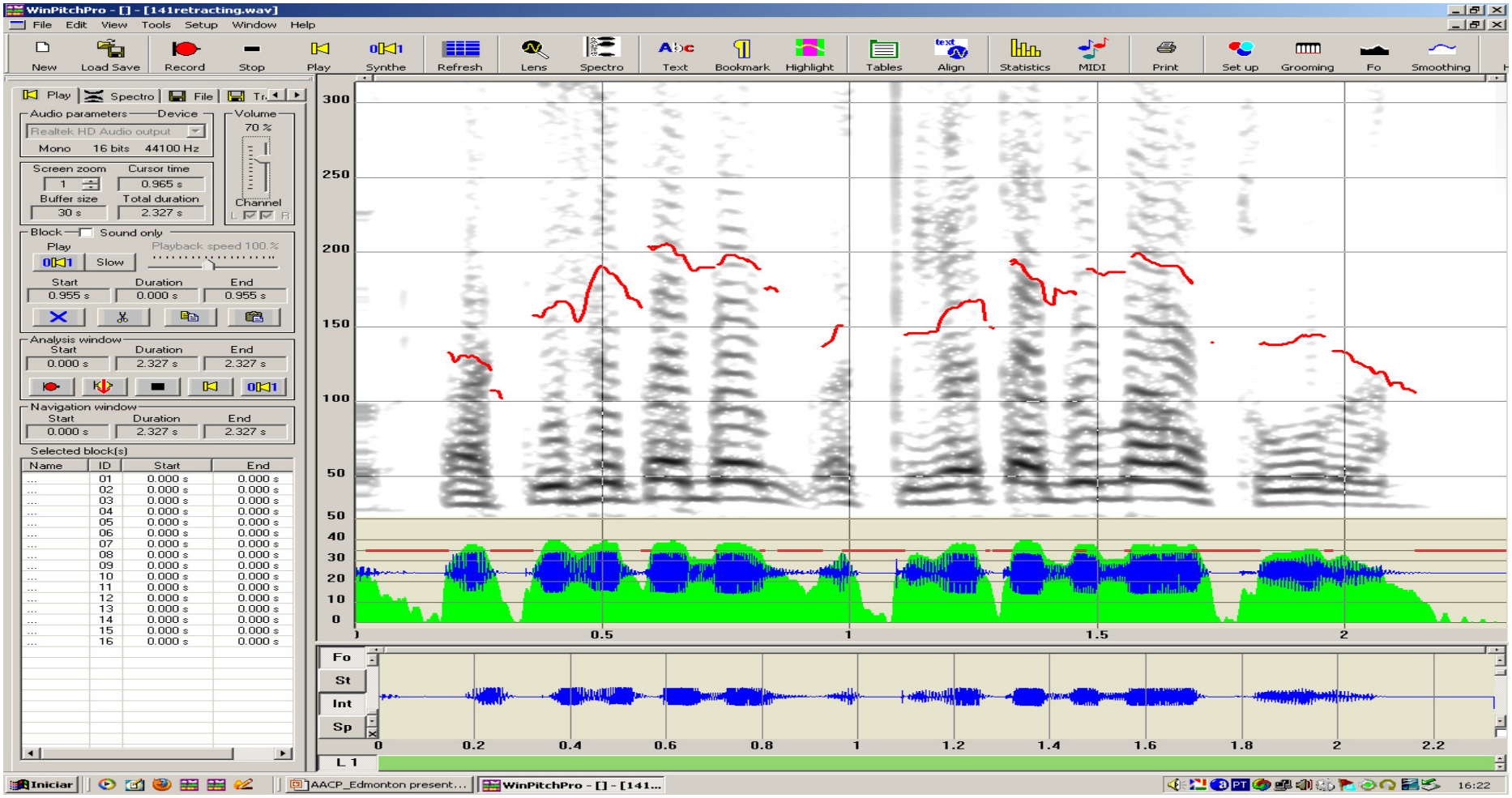
Interrupted utterance (+): when for some reason, the program is not completed.

Tagged complex utterance



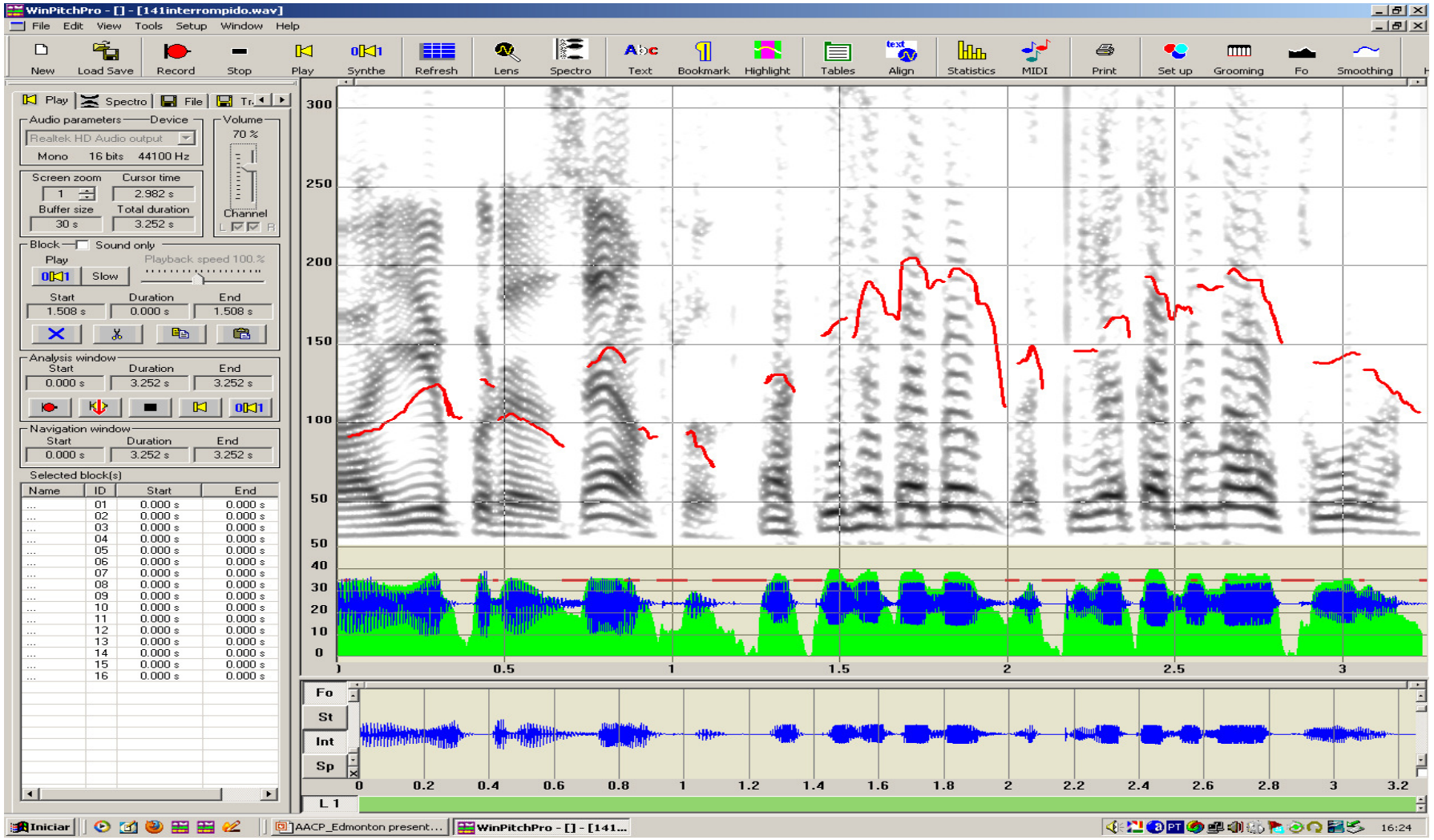
Tudo que é de bom /=TOP= pra gente /APT que a gente tá se sentindo que realmente tá fazendo /=TOP= né /=AUX= e [/] e que tá dando retorno /=APT= a gente continua //=COM=

Retracting



aqui já tá dando [/4] aqui já tá dando a altura //

Interrupted utterance



aí / por exemplo +

Thanks //COM

tommaso.raso@gmail.com
heliana.mello@gmail.com