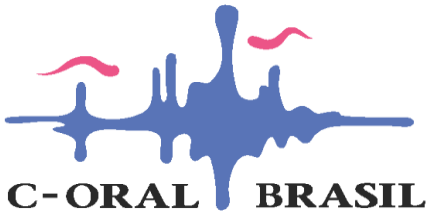
Speech corpora: results and perspectives

Tommaso Raso
UFMG-CNPq-Fapemig
Laboratório de Estudos Linguísticos e Experimentais da Linguagem (LEEL)
tommaso.raso@gmail.com



Summary

1. Speech and corpus linguistics: brief overview of the international and Brazilian context
2. The most relevant methodological aspects and the C-ORAL-BRASIL corpus
   2.1. Spontaneous speech
   2.2. Diaphasy and diastraty
   2.3. Access to audio files and alignment
   2.4. Segmentation
   2.5. Transcriptions
   2.6. Validations
3. Conclusion

Speech and corpus linguistics

What is a corpus*: a computerized data base, with a design for specific objectives (lexicon, phonetics, translation, pragmatics, history of language, etc.), validated and representing a statistical object of study.*

-BNC (1980) 10 MLN words, only recently part of it allows access to sound

- Dutch *corpus* (9 MLN words) (http://lands.let.ru.nl/cgn/doc_English/topics/project/pro_info.htm#intro)

-Santa Barbara *corpus* (600.000 words aligned to sound; prosodic segmentation) (http://www.linguistics.ucsb.edu/research/sbcorpus.html)

- C-ORAL-ROM: comparable corpora

In Brazil

(Mello, in press)

-NURC (1970-1990): only a small part of NURC-RJ allows access to sound http://www.letras.ufrj.br/nurc-rj/

- VARSUL: sample in http://www.varsul.org.br/?modulo=pagina&id=47

- IBORUNA (http://www.iboruna.ibilce.unesp.br/): interviews and chats on the use of northwestern SP state

A new stage in spoken CL

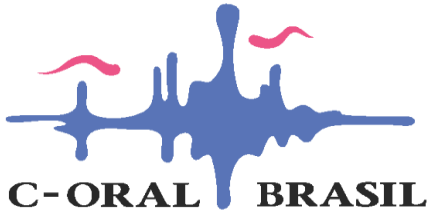
Spontaneous speech (not only chats and interview)

Much more attention to methodological aspects

Reliable data and validations (no validated data, no data)

This is specially true and necessary for speech corpora, whose price (economic and in terms of work) is much higher

The C-ORAL-BRASIL corpus
(Raso & Mello, in press)



- The C-ORAL-ROM project (Cresti & Moneglia, 2005): Spanish, French, Italian and EP (informal and formal: 300.000 words perlanguage)
-C-ORAL-BRASIL informal (210.000 words; 139 texts)

Comparable *Corpora* : to separate what is due to the speech modality and what is due to a specific language

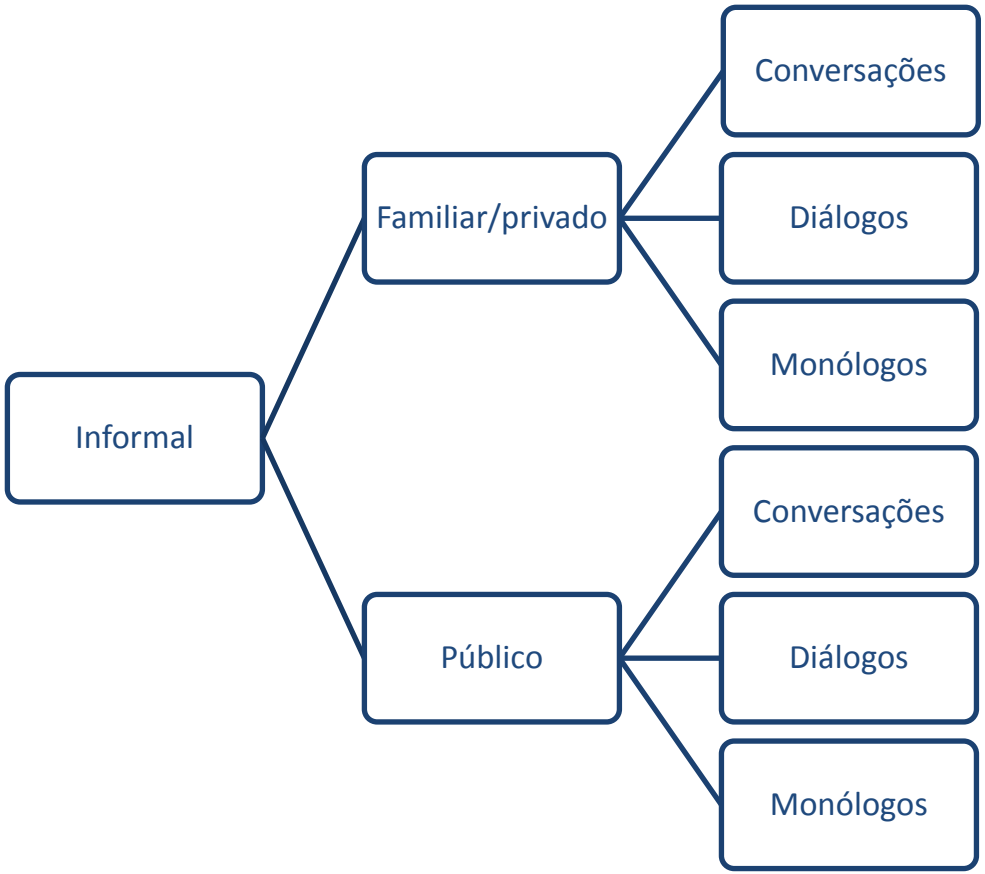
The main methodological aspects of a speech corpus and the C-ORAL-BRASIL corpus

Spontaneous speech: speech that is executed and planned at the same time. It does not execute a previous text (partially or totally planned). It cannot be recorded in laboratory.

Other types of speech: read speech – acted speech – planned speech

(Nencioni, 1983; Cresti, 2000a; Biber, 1988; Blanche-Benveniste *et al.*, 1990; Miller; Weinert, 1998; Givón, 1979; Moneglia, 2005, 2011)

Diaphasic variation



Diaphasic variation and action (with or without movement): illocutions and linguistic structures

people grocery shopping and shoe shopping; construction worker and an engineer at a construction site; driving lesson; people playing pool, soccer and different table games; people cooking or cleaning the kitchen or the house; people working at the computer; a student helping another one with the recorder; driver and passenger talking in a car; waiters waiting at a party; drag-queens putting make up on before a show; a mother telling a story to her child; people telling dramatic moments of their life or explaining their job; jokes; recipes, and many other different situations *(just 14 chats or interviews)*

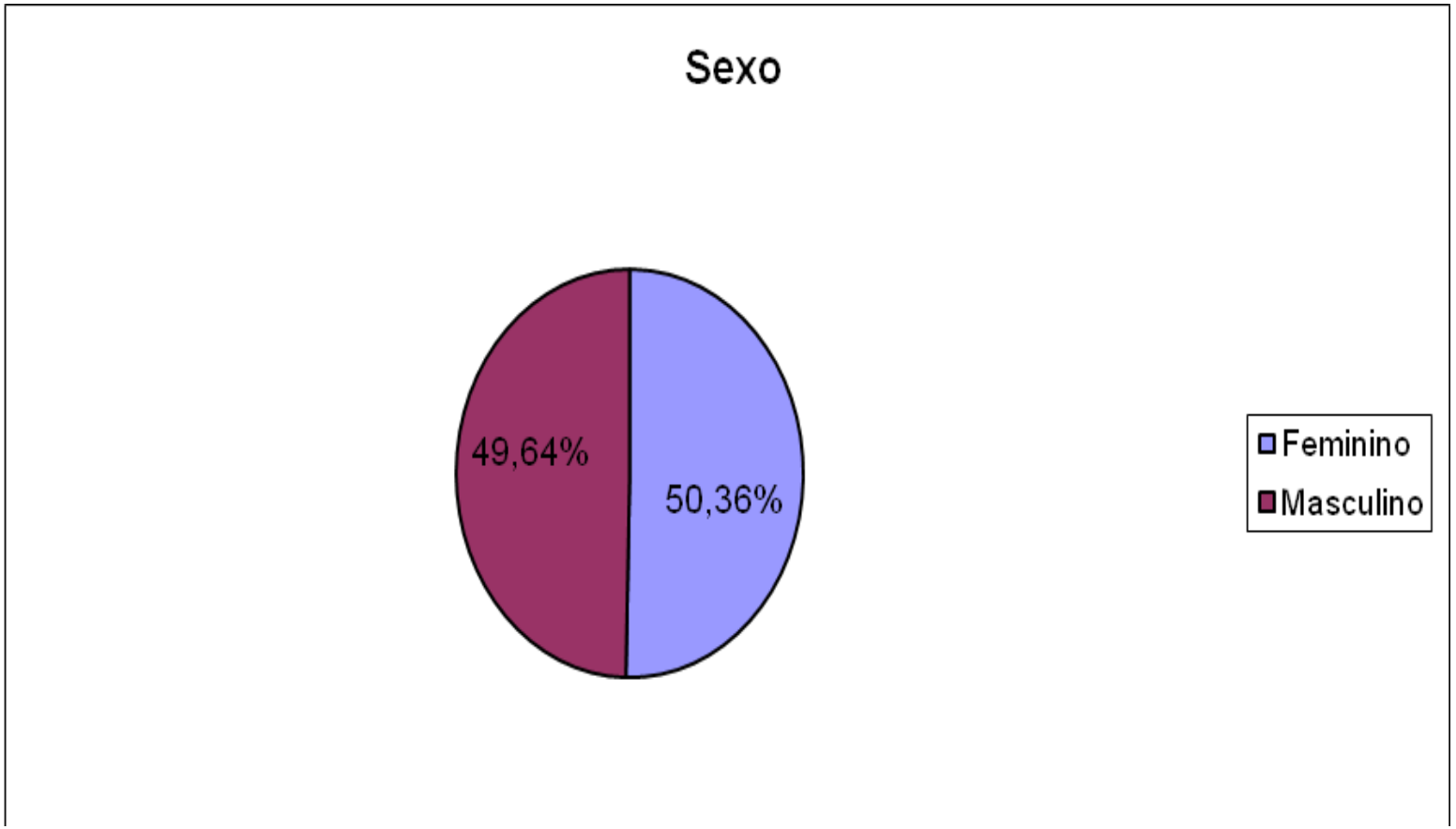
High quality wireless equipment

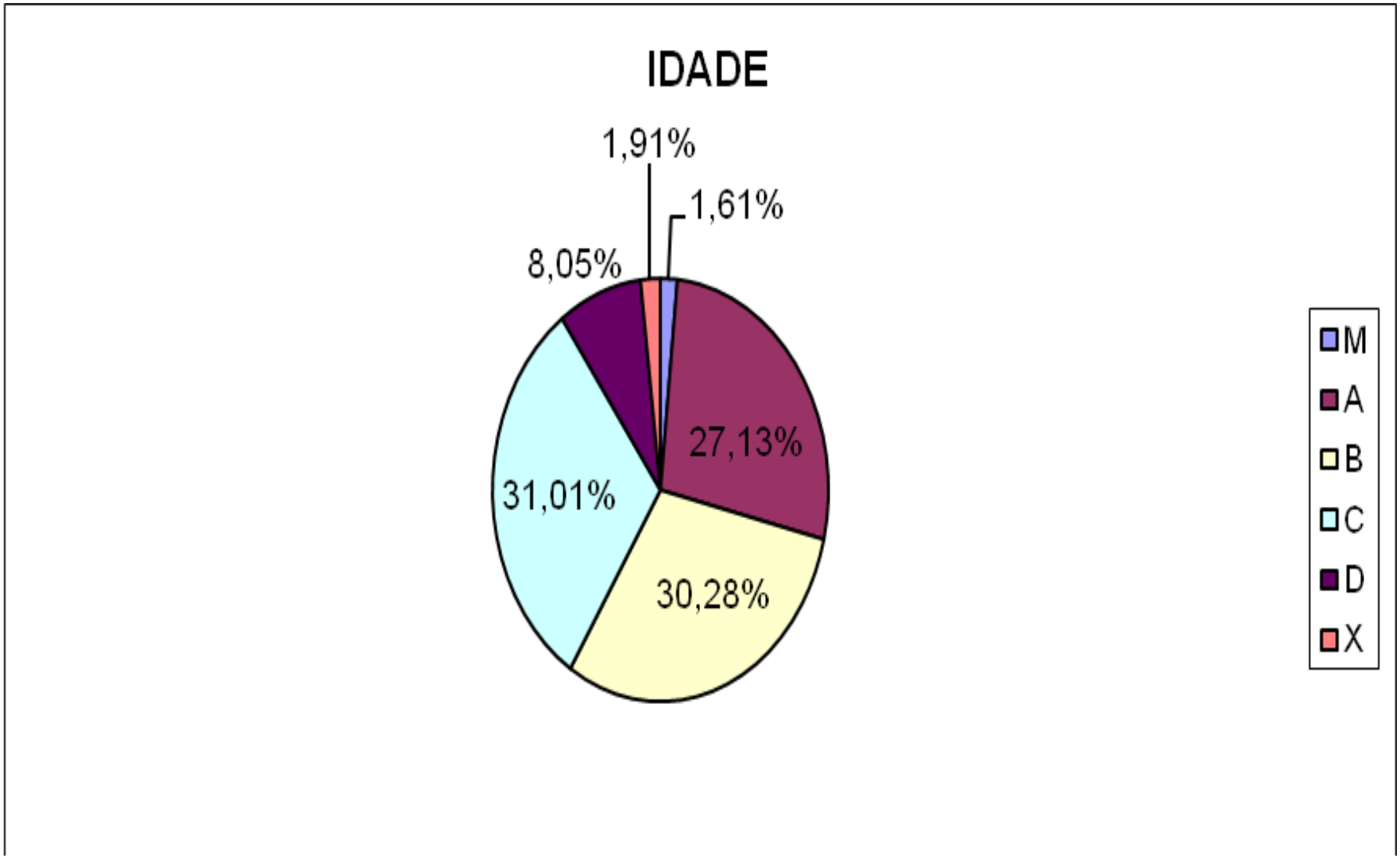
Technical problems → acoustic quality

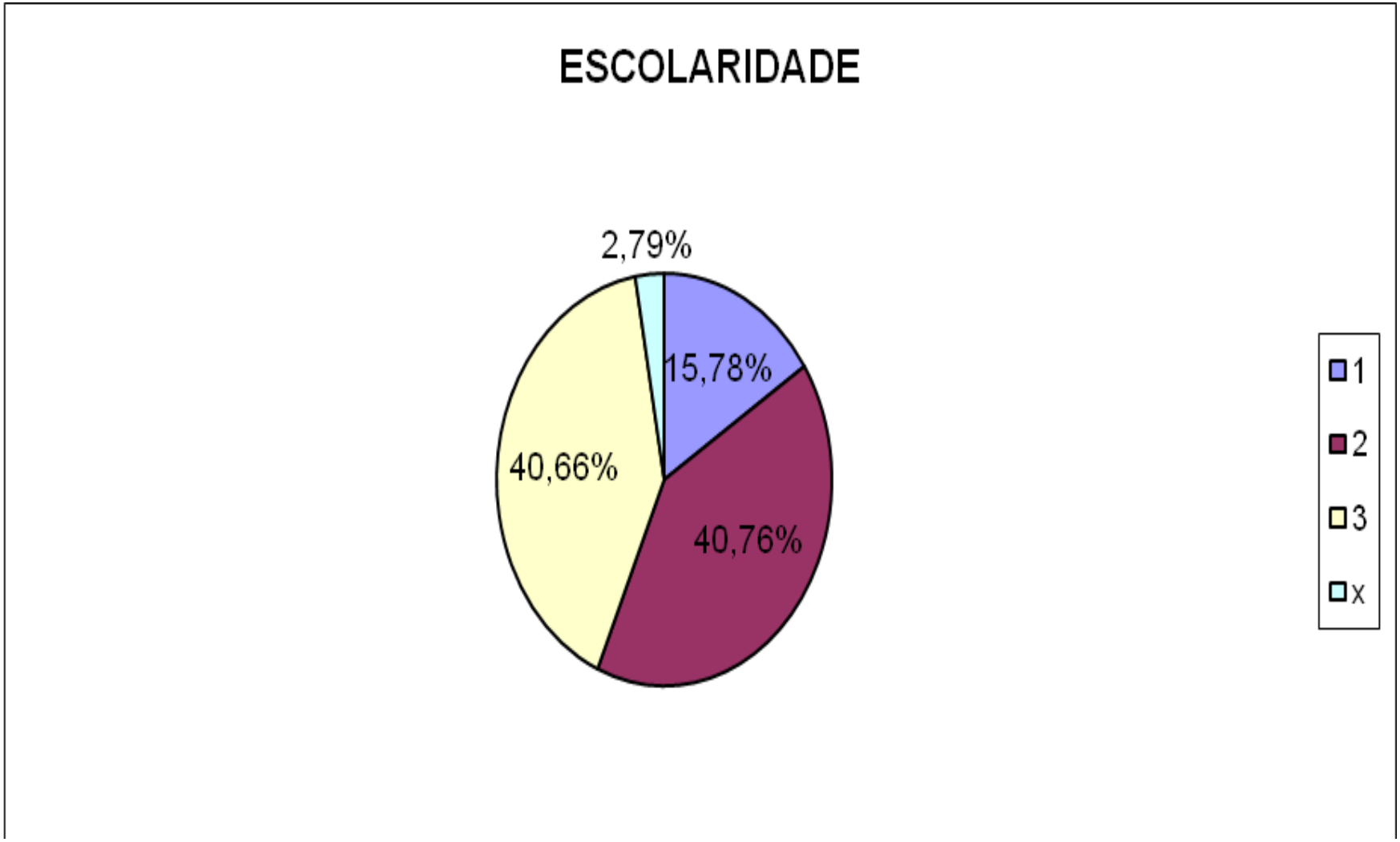
The speaker´s variation

CLUSTER 1	*Speaker´s number*
1 - 247 words (161 speakers)	161
280 - 627 words (81 speakers)	81
649 - 908 words (37 speakers)	37
933 - 1016 words (16 speakers)	16
1134 - 1400 words (26 speakers)	26
1455 - 1663 words (17 speakers)	17
1777 - 1994 words (7 speakers)	7
2140 - 2455 words (10 speakers)	10
2611 - 2901 words (2 speakers)	2
3550 - 3738 words (2 speakers)	2
4211 - 4327 words (2 speakers)	2
6309 words (1 speaker)	1
TOTAL	362

Diaphasy induces diastraty




IDADE
1,91%
1,61%
8,05%
27,13%
31,01%
30,28%
M
A
B
C
D
X


ESCOLARIDADE
2,79%
15,78%
40,66%
40,76%
1
2
3
x

DIATOPY

origin	**speakers**
Belo Horizonte	138
Other cities of Minas Gerais State	89
Other Brazilian states	19
Other countries	2
Unknown	114
Total	**362**

THE REQUIREMENT TO AUDIO ACCESS

Without sound (prosody) speech is not interpretable

Example 1 (*bfammn02*)

*DFL: *e então tinha muito texto do tio Carlos então ele falava ah ele é tio da minha tia* (and so there was uncle Carlos text so he said ah he's my aunt's uncle)

Example 2 (*bpubdl01*)

*PAU: *não tá dando a altura daquele que a Isa marcou lá né (it does not [it] have [has] the height that Isa signed there)*

THE ALIGNMENT

The access to the sound is not sufficient

It is necessary to have the text aligned to the sound

www.winpitch.com (Ph. Martin)

THE SEGMENTATION

Reference unit for speech

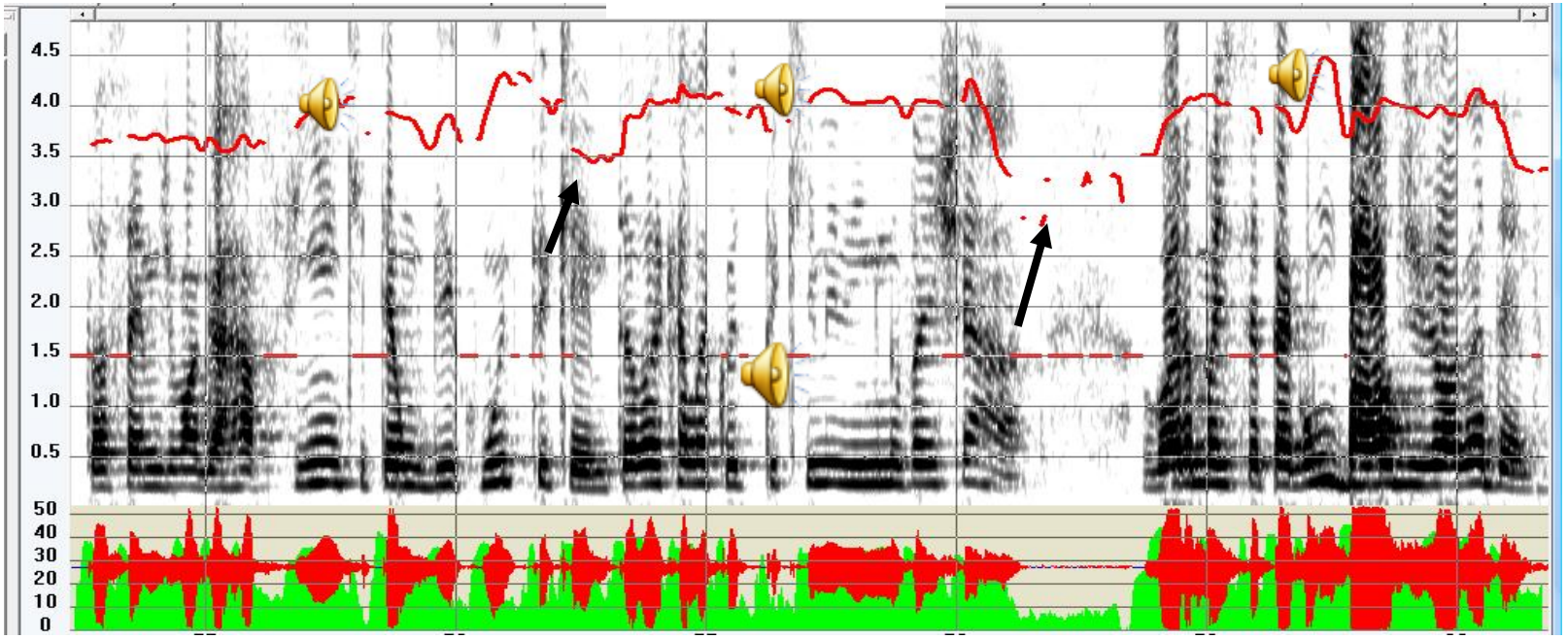
The utterance: the smallest part of the speech flow pragmatically interpretable

It can be identified by a prosodic terminal break (//) (not for pause)

Examples

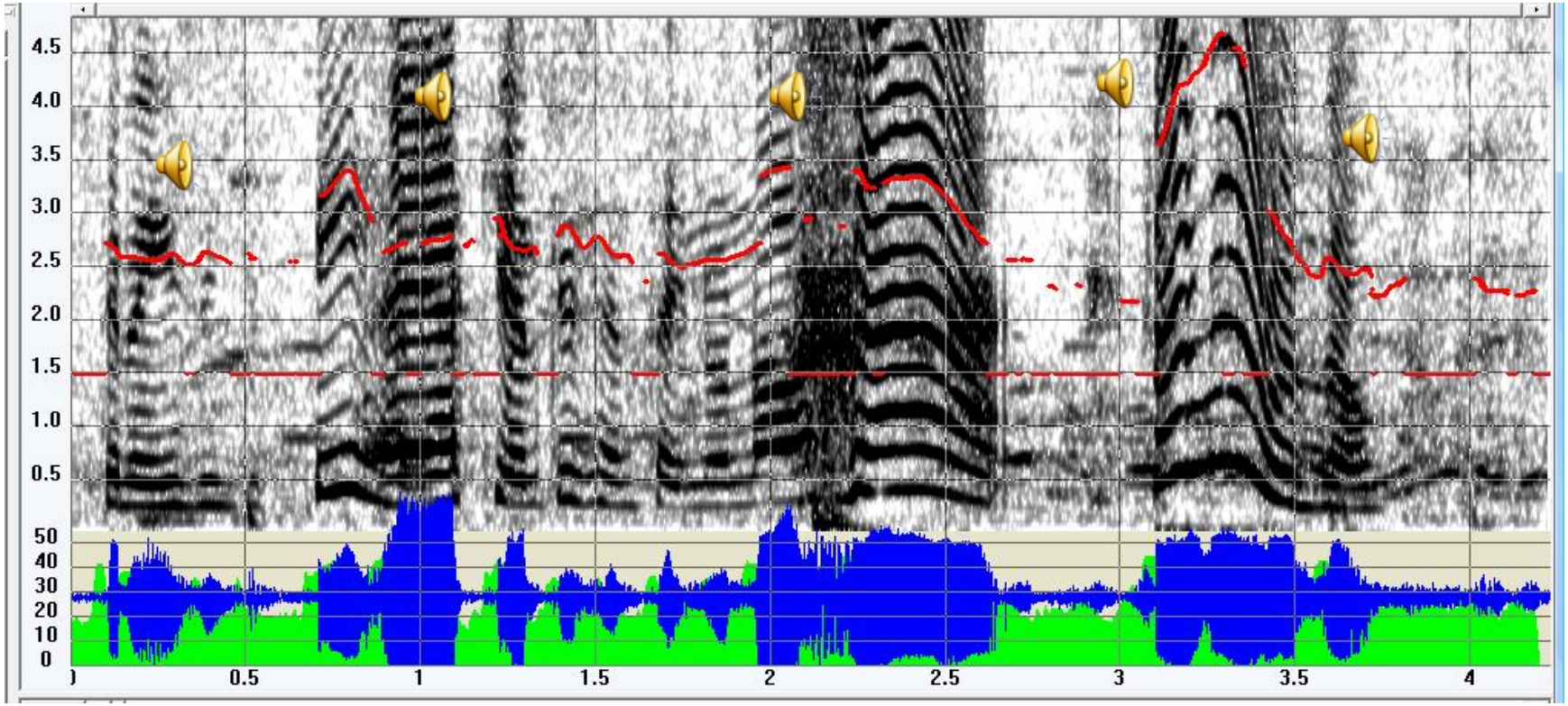
tá saindo de uma garrafinha que tem um bico muito pequeno // então daquela coisa pequeninim nũ vai encher rápido // agora imagina cê pega um balde e joga dentro // It's coming out from a little bottle with a very small neck // so that little thing can't fill it quickly // now you imagine you fill it with a full bucket //

C-ORAL BRASIL



REN: trenzim que espirra // a little thing that sneeze //
FLA: é // aquele que a gente tem no norte // yeah // that one we have in the North //
REN: ah // cês usam // ah // you use (it) //

C-ORAL BRASIL



Tone units segmentations

Simple utterance: made by only one TU. In this case the TU is a COM IU. The COM is the necessary and sufficient IU to build na utterance, as it carries the illocutionary force (and the pragmatic and prosodic autonomy)

Complex utterance: made by at least the COM and one or more TU, with functions different from that of carrying the illocutionary force (/).

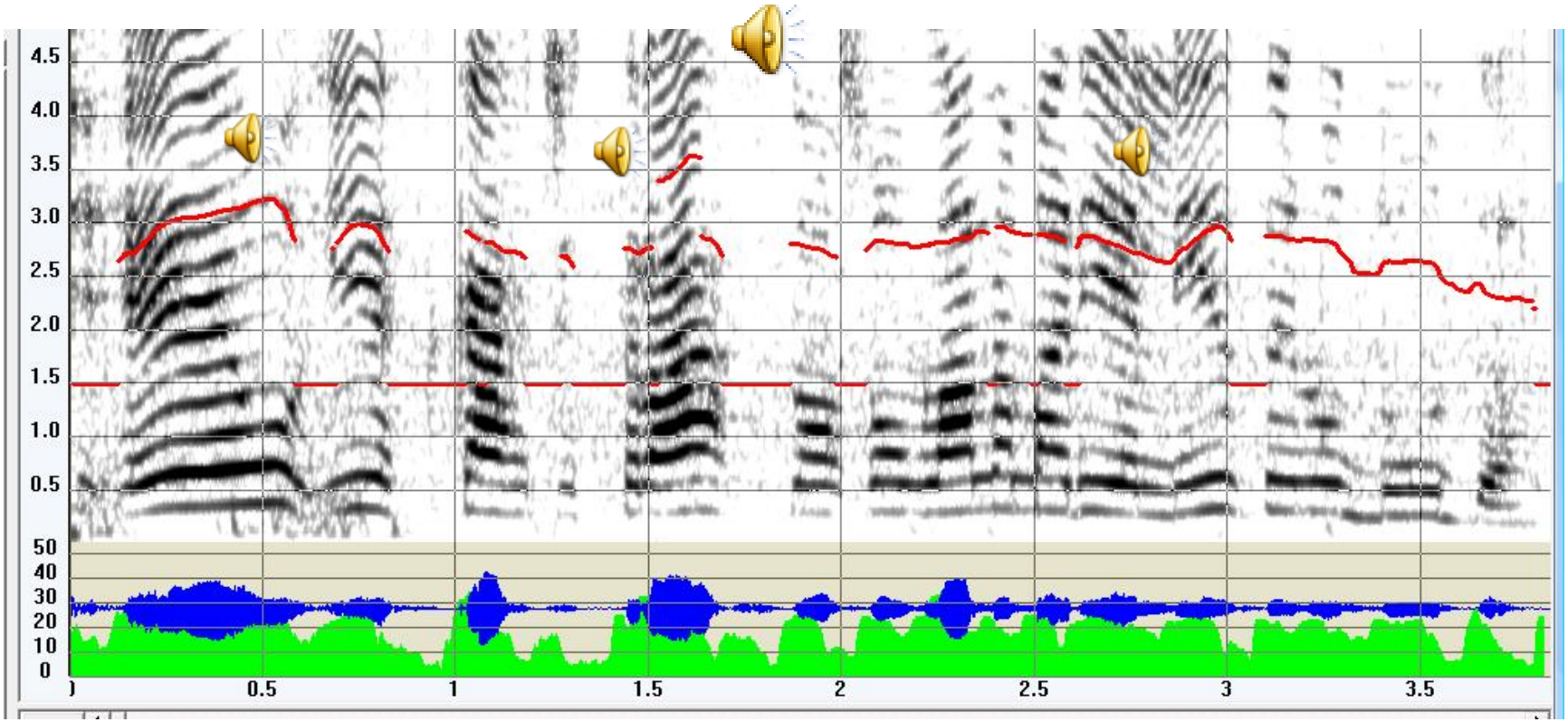
C-ORAL BRASIL

Examples

LUZ: eu /=TOP= e esse carro de trás /=TOP= nós vamo lá na Maria Eliza e no Duda //=COM= (me / and the car behind / we are going to Maria Elisa's and Duda's)

C-ORAL BRASIL

Only the COM is autonomous and interpretable in isolation



The transcriptions

The transcription's goals are to be:

1. Readable
2. Preserve typical speech phenomena (not normalizing them according with the orthographic idealization of writing)
3. Computable
4. Validable

SOME OF THE PHENOMENA PRESERVED IN THE TRANSCRIPION (Mello & Raso, 2009)

- 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.

VALIDATIONS

Segmentation validation

The transcribers interrater agreement shows a kappa of 0,86 (RASO-MITTMANN 2009).

Transcriptions validations

Also the segmental part of transcriptions was validated: 0,81% mistakes; 0,57% mistakes for non-orthographic criteria. The worst phenomenon shows 3,25% of mistakes.

Important basis for any morfosyntactic research with statystical and computables technique

SPEECH AND CORPORA

BELO HORIZONTE (UFMG), february 29th – march 2nd 2012

www.letras.ufmg.br/gscp2012

V. Aubergé (Univ. Grenoble) - P. Barbosa (Unicamp) - P. Bertinetto (Scuola Normale Superiore) - D. Biber (Univ. of Northern Arizona) – E. Bick (Univ. of Southern Denmark) – E. Cresti e M. Moneglia (Univ. di Firenze) -Ph. Martin (Univ. Paris Diderot) - J. Moraes (UFRJ) - K. Scherer (Univ. Genève)