

DESIGN DOC.

Specifications on the C-ORAL-BRASIL Informal Corpus¹

Tommaso Raso

(Universidade Federal de Minas Gerais, Belo Horizonte)

Tommaso Raso
Scientific Manager of the C-ORAL-BRASIL project
LEEL
Faculdade de Letras
Universidade Federal de Minas Gerais
Av. Antônio Carlos 6627
Pampulha – 31270-901
Belo Horizonte
Brazil
e-mail tommaso.raso@gmail.com
web <http://www.c-oral-brasil.org>

¹ The parser description and evaluation are due to Eckhard Bick, developer of the parser.

<i>I Introduction</i>	5
<i>II General Information</i>	7
1. Contact person	7
2. Distribution media	7
3. Content	7
4. Format of speech and label file	8
5. Layout of the disk-file system	8
6. Hardware, Software and Recording Platforms	10
6.1 Hardware	10
6.2 Recording Equipment and Software	10
7. Number of recordings and general corpus data	11
1. Corpus Design	12
1.2 Sampling Parameters	12
1.3. Sampling strategy and Corpus design	12
1.4 Comparability	14
2. Filenames conventions	16
3. Annotation information	16
3.1 Meta-Data	16
3.1.2 Rules for the <i>Participant</i> field	18
3.1.3 Rules for the <i>Class</i> field	19
3.1.4 Rules for the <i>Situation</i> field	19
3.1.5 Rules for marking <i>Acoustic quality</i>	19
3.1.6 Quality assurance on metadata format	20
3.1.7 Statistics from C-ORAL-BRASIL metadata	20
3.1.7.1 Number of speakers	20
3.1.7.2 Distribution of speakers per geographical origin	20
3.1.7.3 Completeness of speakers features in the metadata records	22
3.1.7.4 Completeness of session metadata	23
3.2. Transcription and Dialogue representation	23
3.2.1 Basic Concepts for dialogue representation	23
3.2.2. Turn representation.	24
3.2.3 Utterance representation.	24
3.2.4. Word representation.	24
3.2.5 Transcription	24
3.2.6 Overlapping	25
3.2.7 Cross-over dialogue convention	25
3.2.7.1 Overlapping and cross-over dialogue	26
3.2.7.2 Intersection of turns	26
3.2.8 Transcription conventions for segmental features	26
3.2.8.1. Other transcription conventions	26
3.2.8.1.1 <i>Hesitations and interrupted words</i>	26
3.2.8.1.2 <i>Onomatopoeias</i>	26
3.2.8.1.3 <i>Interjections and exclamations</i>	26
3.2.8.1.4 <i>Abbreviations and acronyms</i>	27
3.2.8.1.5 <i>Numerals</i>	27
3.2.8.1.6 <i>Foreign words and mispronunciations</i>	27
3.2.8.1.7 <i>Forms that are transcribed following the standard orthography even if pronounced differently</i>	27
3.2.8.3. Non-understandable words	29
3.2.8.4 Paralinguistic elements	29
3.2.8.5. Fragments	29

3.2.8.6. Interjections	30
3.2.8.7. Non-standard words	30
3.2.8.8. Non-transcribed words	30
3.2.8.9. Non-transcribed audio signal	30
3.2.9 Quality assurance on format and orthographic transcription	30
3.2.10 Transcription validation	31
3.2.10.1 Initial validation	31
3.2.10.2 Final validation	32
3.3. Prosodic annotation scheme	33
3.3.1 Principles	33
3.3.2 Concepts	33
3.3.3 Theoretical background	33
3.3.4 Conventions for prosodic tagging in the transcripts: types of prosodic breaks	34
3.3.4.1 Terminal breaks (utterance limit)	34
3.3.4.2 Non-terminal breaks	34
3.3.5 Fragmentation phenomena	35
3.3.5.1 Interruptions	35
3.3.5.2 Retracting and/or restart and/or false start(s)	35
3.3.5.3. Retracting/interruption ambiguity	36
3.3.6 Summary of prosodic break types	36
3.4 Procedures for prosodig tagging	36
3.5 Quality assurance on prosodic tagging	36
3.5.1 Trancier training procedures	36
3.5.2 Prosodic segmentation validation	37
3.5.2.1 Previous validation	37
3.5.2.1.1 Group 1	37
3.5.2.1.2 Group 2	40
3.5.2.2 Group 1 reevaluation and final validation	45
3.6. Alignment	46
3.6. 1 Annotation procedure	47
3.6.2. Prosodic tagging and the alignment unit	47
3.6.3 Quality assurance on the alignment	47
3.6.4 WinPitch Pro	47
3.7. PoS tagging and lemmatization	48
3.7.1 Tag sets	48
3.7.2 Automatic PoS tagging: tool and evaluation	62
<i>The Morphosyntactic Tagging of the C-ORAL-Brasil Corpus</i>	62
3.7.2.1 Introduction	62
3.8 Tagging format and categories	63
3.8.1 Two-level annotation	63
3.8.2 Morphosyntactic tag fields	65
3.8.2.1 Word form	65
3.8.2.2 Lemma / Base form	65
3.8.2.3 Secondary tags	65
3.8.2.4 Part of speech	66
3.8.2.5 Morphology / inflexion	66
3.8.2.6 Syntactic function	67
3.8.2.7 Dependency links and syntactic trees	68
3.9 The tool	69
3.10 Tokenization	71
3.10.1 Multi-word expressions	71
3.11 Lexical and orthographic normalization	74

3.12	Syntactic segmentation	77
3.13	Evaluation	78
REFERENCES		81
Appendixes		83
<i>Appendix 1- Typical examples of prosodic breaks types in Brazilian Portuguese</i>		84
<i>Appendix 2 - Tagsets used for PoS tagging in Brazilian Portuguese. Detailed tables and comparison table</i>		85
<i>Appendix 3 Orthographic transcription conventions in Brazilian Portuguese</i>		90
<i>Brazilian transcription conventions</i>		90
<i>Appendix 4 C-ORAL-BRASIL Prosodic Tagging Evaluation Report</i>		120
<i>Appendix 5 C-ORAL-BRASIL Transcription validation report</i>		129

I Introduction

The C-ORAL-BRASIL resource provides a Brazilian Portuguese informal spontaneous speech corpus comparable with the corpora of the C-ORAL-ROM Project for French, Italian, European Portuguese and Spanish (Cresti & Moneglia, 2005). These specifications, with the agreement of the C-ORAL-ROM author, have, with a few differences, the same structure of the C-ORAL-ROM specifications. It is important to say that C-ORAL-BRASIL has not completed the formal part of the corpus so far, nevertheless it shows some methodological and technical developments in regard to C-ORAL-ROM.

C-ORAL-BRASIL is a project associated with the Laboratório de Estudos Empíricos e Experimentais da Linguagem (LEEL) of the Faculdade de Letras of the Universidade Federal de Minas Gerais (UFMG), Brazil. The main goal of the project is to offer a spontaneous speech corpus of Brazilian Portuguese for the study not only of lexis and morphosyntax, but also of pragmatic categories, such as information structure and illocution. The project was financed by the Fundação de Amparo à Pesquisa do Estado de Minas Gerais (Fapemig), by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), and by the Universidade Federal de Minas Gerais (UFMG). The project was executed in the framework of an institutional agreement of collaboration between the Università di Firenze (Italy) and the Universidade Federal de Minas Gerais (Brazil), coordinated by Emanuela Cresti and Tommaso Raso.

C-ORAL-BRASIL consists of 139 spoken texts and 21:08:52 hours of speech (208.130 words). The resource aims to represent the variety of speech acts performed in everyday language and to enable the induction of prosodic and syntactic structures in Brazilian Portuguese, from a quantitative and qualitative point of view. More specifically, the representation of significant variations found in spontaneous speech performances in natural environments allow the use of C-ORAL-BRASIL for comparable spoken language modeling of the main Romance languages, both for linguistic studies and language technology purposes. The resource can also be used for testing speech recognition tools in critical contexts.

The recording conditions and the acoustic quality of the sessions collected in C-ORAL-BRASIL are variable. The speech files from the acoustic database are defined on a quality scale (recording, volume, voice overlapping and noise) and are comparable with respect to it. The quality scale extends from the highest level of clarity of the voice signal (A) to low levels of acoustic quality (C). The quality is gauged spectrographically and is always annotated in the metadata of each session. This scale is not comparable with that used in C-ORAL-ROM, as the average acoustic quality of C-ORAL-BRASIL is much higher.

The C-ORAL-ROM databases are anonymous. All speech segments that may have offended the user for decency reasons have been erased and substituted with a beep in the audio signal. Speakers authorized each provider for the use of the recorded data to all ends foreseen in the C-ORAL-ROM project, including publication and language technology applications. The authorization models are available at the Laboratório de Estudos Empíricos e Experimentais da Linguagem (LEEL) at the Faculty of Letras of the Universidade Feral de Minas Gerais.

In order to ensure a significant representation of the spontaneous speech universe, the corpus design of the resource foresees recording in natural environment in a variety of different contexts. The contextual variation is controlled on the basis of a strictly defined set of parameters whose significance has been recognized in the linguistic tradition. As a consequence of this sampling strategy, the C-ORAL-BRASIL and the C-ORAL-ROM resources are comparable as far as they fit in the same corpus design scheme.

Each recorded session is stored in wav files (Windows PCM, 22.050 Hz, 16 bit) and is delivered in a multimedia corpus with the following main annotations:

- a. Session metadata;
- b. The orthographic transcription, in CHAT format (MacWhinney 1994)², enriched by the tagging of terminal and non terminal prosodic breaks, in .txt files
- c. The text-to-speech synchronization, based on the alignment to the acoustic source of each transcribed utterance, in .xml files.

This resource is stored in one DVD and can be fully exploited with WinPitch software (www.winpitch.com). WinPitch allows the direct and simultaneous exploitation of the acoustic and textual information by loading the .xml alignment files, compiled in accordance with the C-ORAL-ROM DTD alignment.

Metadata are defined following an explicit set of rules and contain essential information regarding the *speakers*, the *recording situation*, the *acoustic quality*, the *source*, and the *content* of each session and ensure a clear identification of the various speech types documented in the resource (see. § 3.1 in Chapter II)

The corpora are orthographically transcribed in standard textual format (CHAT format; MacWhinney 1994) with the representation of the main dialogue characters; that is, *speaker's turns*, the occurring *non linguistic* and *paralinguistic events*, *prosodic breaks* and the segmentation of the speech flow into discrete speech events. (See § 3.2 in Chapter III)

In conformity with C-ORAL-ROM's implementation, the textual string is divided into *utterances* following the annotation of perceptively relevant prosodic breaks, which are discriminated in the speech flow through perceptive judgments. (See §3.3 in Chapter III).

The annotated transcripts are aligned to the acoustic counterpart through *WinPitchPro*. Segments deriving from the alignment are defined on independent layers, with automatic generation of the corresponding database.

This multimedia storage ensures a natural and meaningful text/sound correspondence that is one of the main added values of the C-ORAL-ROM. C-ORAL-BRASIL is, therefore, comparable with C-ORAL-ROM in respect to this, too. Each utterance is aligned to its acoustic counterpart, generating the database of all the utterances in the resource (34,167) (see § 3.6 in Chapter III).

Besides text-to-speech and speech-to-text alignment, *WinPitchPro* allows an easy and efficient acoustic analysis of speech, as regards real-time fundamental frequency tracking, spectrographic display, re-synthesis after editing of prosodic parameters, etc... *WinPitchPro* is able to export the alignment in *Praat*.

The multimedia C-ORAL-ROM resource is integrated with additional label files in various formats, which ensure a multitask exploitation of the resource:

- TXT files with the resource metadata in CHAT format
- Textual resource without alignment information in TXT files
- Textual resource with automatic Part of Speech (PoS) and lemma tagging of each form
- Textual resource in XML format according with the C-ORAL-ROM DTD

² <http://chilides.psy.cmu.edu/manuals/CHAT.pdf>

The project ensures maximum accuracy in the transcripts, which have been compiled by PhD, MA and senior Undergraduate students in linguistics. The original transcripts have been revised by at least four transcribers. The orthographic accuracy of transcripts has been checked manually and through automatic PoS tagging.

The reliability of the prosodic tagging has been validated through an interrater agreement test: Kappa statistic 0.86 (see. the evaluation report in Appendix IV).

The reliability of the orthographic transcriptions has been validated (see the evaluation report in Appendix III).

The level of accuracy of the automatic PoS tagging has been evaluated by its author, Eckhard Bick, and reported in §3.13. The percentage of exact recognition is always over 94,9%.

II General Information

1. Contact person

Prof. Tommaso Raso
Coordinator of the C-ORAL-BRASIL project
Faculdade de Letras
Universidade Federal de Minas Gerais
Av. Antônio Carlos 6627
Pampulha – 31270-901
Belo Horizonte
Brazil
phone: +55 031 34096091/34096094
e-mail: tommaso.raso@gmail.com
web: <http://www.c-oral-brasil.org>

2. Distribution media

The C-ORAL-BRASIL resource is distributed in DVD-DL.

3. Content

The C-ORAL-BRASIL resource of spontaneous speech comprises the following components:

- a) Multimedia corpus;
 - b) Textual corpus;
 - c) Appendix;
 - b) PoSTagged corpus;
 - c) Book;
 - d) Readme&Copyright file.
-
- a) Multimedia corpus comprises all the corpus files in three different formats: wav format for the sound; rtf format for the text; xml files for the alignment;
 - b) Textual corpus comprises all the corpus files in txt format;
 - c) Appendix comprises the frequency lists, the corpus measurements and the speakers statistics, the metadata of all corpus files, the prosodic breaks types, and the resource specifications;

- d) PoStagged corpus comprises the tag set and all corpus PoS tagged files in four versions: full PoStagged version in txt format; full PoStagged version in xml format; simplified PoStagged version in txt format; simplified PoStagged version in xml format;
- e) The book, in pdf format, has the sound linked to all the text examples.

4. Format of speech and label file

For each spontaneous speech recording session, the following is delivered into folders of the multimedia corpus

1. Speech files: uncompressed files (Windows PCM: 22,050 Hz; 16 bit) with “.wav” extension
2. Transcripts in CHAT format³ enriched by the annotation of terminal and non terminal prosodic breaks (Moneglia & Cresti 1997) and the alignment information, in plain text files with “.rtf” extension
3. The text-to-speech alignment files: XML file in WIN PITCH PRO format with “.xml” extension.
4. Metadata file for each recorded session corresponding to a wav file, an xml file and an rtf file. The metadata files are with “.txt” extension
5. DTD of the WinPitchPro alignment format (alignment.dtd)
6. The C-ORAL-BRASIL transcription of each session with Part of Speech annotation and Lemma annotation for each form in plain text files with “_PoS.txt” extension

In addition, the following files are delivered for the resource:

7. Tag set adopted in plain text files (tagset_brazilianportuguese)
8. Frequency lists of lemmas and Frequency lists of forms in plain text files
9. Measurements of the Language values recorded in each text: in the Excel files “measurements_language.xls”
10. A set of excel files containing statistics on the metadata: a) metadata_session.xls: statistics regarding the completeness of session metadata; b) participants_records.xls: statistics regarding the completeness of the main speaker metadata; c) list of speakers recorded in the corpus, in anonymous form, with their main metadata including geographical origin, sex and age.
11. A set of multimedia samples referred to in the resource documentation.

In the DVD there is also a book about the project

Standard character set used for transcription and annotation: ANSI (the transcriptions contain the special character “ũ” encoded with the rtf Unicode escape sequence for non ANSI characters: “\u361?”)

5. Layout of the disk-file system

The language collection has a folder structure, which mirrors the C-ORAL-BRASIL corpus design⁴:

/<Category>/<Context>/<Domain>

where:

Category	INFORMAL
Context	family_private public
Domain	monologues dialogues conversations

³ <http://childes.psy.cmu.edu/manuals/CHAT.pdf>

⁴ The Category folder is due to the fact that the formal part is also foreseen.

The following is the directory structure of the language database in the C-ORAL-BRASIL multimedia resource:

```
INFORMAL/  
INFORMAL/ family_private  
INFORMAL/ family_private / monologues  
INFORMAL/ family_private / dialogues  
INFORMAL/ family_private / conversations  
INFORMAL/public  
INFORMAL/public/dialogues  
INFORMAL/public/ conversations  
INFORMAL/public /monologues
```

The DVD contains a set of Appendixes to the multimedia and the textual corpus (Textual resources in various formats, and additional documentation). The content is structured into folders as follows:

```
\Appendix  
\Specifications  
Frequency Lists  
Measurements  
Prosodic-Breaks-Types  
Metadata  
  
\PoS-Tagged-Corpus  
\PoStagged_files.cg  
PoStagged_files.xml  
SimplifiedPoStagged_files.txt  
SimplifiedPoStagged_files.xml  
PoS-Tagset  
  
\Book
```

In addition to the previous structures, the following directories are used to store those files which are not part of the Data Base:

\ (root) README©RIGHT.TXT file containing a short description of the database and the files, and copyright and use conditions

6. Hardware, Software and Recording Platforms

6.1 Hardware

Computers:

Processor: Intel Pentium IV 1,8 GHz or higher

RAM: 1 GB or more

Hard Disk: SATA 100GB or higher

6.2 Recording Equipment and Software

RECORDING EQUIPMENT

Recorder

Marantz PMD660 Professional Solid State Recorder

Unidirectional wireless microphones

Unidirectional microphones **Sennheiser ME 4** clip-on (cardioid)

Cable **Sennheiser CL100** (connectors XLR and mini jack 1/8")

Receiver

- **Sennheiser EM 100 G2 A**
- **Sennheiser EK 100 G2 A**
- **Sennheiser EK 100 G3**

Transmitter

- **Transmitter Sennheiser SK 100 G2 A**
- **Transmitter Sennheiser SK 100 G3**

Omnidirectional Microphones

omnidirectional microphone **Sennheiser MD 421-II 4**

omnidirectional microphone **Shure PG58-XLR**

Mixer

Behringer XEXYX 1222 FX:

Software

Sound Editor: Audacity®1.2.6

Text Editor: Microsoft® Word© 2000 or higher

Text to Speech Alignment Editor: WinPitchPro – Pitch France©

PoS tagging: Adapted version of PALAVRAS Parser

7. Number of recordings and general corpus data

Audio files of the recorded sessions and the corresponding transcription files are in one to one correspondence. The following is the general table of the main values recorded in the C-ORAL-BRASIL informal multimedia corpus, compared with C-ORAL-ROM informal corpora. However C-ORAL-ROM does not differentiate informal and formal speakers; so numbers of the categories with asterisk refer to both informal and formal corpora. In C-ORAL-BRASIL some calculations have been done both with the macro used for C-ORAL-ROM and with an R language script. The results have small differences, but R results are more precise. We give them into parenthesis.

	wav files	GB	Duration	Utterances	Words	Speakers*	Male*	Female*
BP	139	5,6	21:08:52	34.151 (34.167)	208.130	362⁵	158	203
French	98	1,83	12.09.54	10517	152.385	305	154	150
Italian	92	2,60	16.50.49	23805	154.967	451	276	175
EP	86	2,37	14.56.31	21949	165.436	261	144	117
Spanish	89	2,19	14.36.21	21618	168.868	410	247	163

*The number of speakers and their sex refers to both informal and formal sections in the four languages of C-ORAL-ROM

8. Acoustic quality

As C-ORAL-ROM, C-ORAL-BRASIL is oriented towards the collection of corpora in natural environment, despite the fact that this necessarily causes a lower acoustic quality of the resource.

The following are the requirements for the acoustic format and for the recording apparatus:

Format: mono and stereo wav files (Windows PCM), Sampling frequency: 22050Hz, 16 bit⁶

Recording and storing process for new recordings:

a) *dialogues and monologues:* stereo or mono digital recording (44.100Hz) with unidirectional Microphones, saved in mono or stereo .wav files (Windows PCM, 22050Hz, 16 bit);

b) *conversations with more than two participants:* mono recording with omnidirectional microphone in mono .wav files or stereo recording and up to eight channel stereo recording with unidirectional microphones mixed in two channel stereo files (44.100Hz) converted in (Windows PCM, 22050Hz, 16 bit.

All the recordings are digital. The speech files of the acoustic database are defined on a quality scale (volume, voice, overlapping and noise). The quality scale extends from the highest level of clarity of the voice signal to low levels of acoustic quality. The C-ORAL-BRASIL quality scale is not comparable to the one of C-ORAL-ROM, because C-ORAL-BRASIL has a much better average quality, due to new technologies, to the use of a mixer for conversation recordings and to the fact that in C-ORAL-ROM older resources were also exploited, especially for Italian and European Portuguese. The use of a mixer for conversations was a very important innovation in reference to the techniques used in C-ORAL-ROM. Actually, only a very small number of the recordings with omnidirectional microphone reached an acceptable acoustic quality. The use of the mixer, which allowed the use of monodirectional microphones for all the participant speakers, made the average quality of conversations much better.

The quality is gauged spectrographically. Sessions in which F0 analysis is not significant are excluded from the sampling. The acoustic quality of each recording and the most relevant data on the recording condition are always recorded in the metadata of each text.

⁵ The sex of one informant is unknown. This informant says just one word.

⁶ See footnote 7

III C-ORAL-BRASIL corpus

1. Corpus Design

1.2 Sampling Parameters

Spontaneous speech events are those *communication events in which the programming of speech happens simultaneously to its execution by the speaker*; i.e. the speech event is non-scripted or only partially scripted. C-ORAL-BRASIL resource offers a representation of the spontaneous speech universe in Brazilian Portuguese with regard to the following main parameters, which define speech variation:

A. Communication context

B. Language register

C. Speaker

A. The communication context of a speech event is defined by the following features, each specified by a closed vocabulary:

Channel: the means by which the signal transmission is achieved.

Face to face communication: speech event among participants in the same unity of space and time with reciprocal direct multimodal perception and interaction;

Structure of the communication event: role and nature of the participants in the speech event.

Monologue: speech event with only one intervenient performing a main communication task⁷

Dialogue: speech event with two intervenients

Conversation: speech event with more than two intervenients

Social context: organization level of the society to which the speech event belongs

Family/private: speech event within the family, or private social context

Public: speech event within a public social context

B. Language register

Informal: unscripted low variety of language, used for everyday interactive purposes

C. Speaker: the main qualities of the speaker that may influence their speech production

Sex: sex of the speaker

Age: speaker's age

Education: speaker's schooling degree

Occupation; speaker's employment field

Geographical origin: speaker's place of origin

1.3. Sampling strategy and Corpus design

The corpus design of the C-ORAL-BRASIL resource is the result of the application of two criteria:

- a sampling strategy, which defines how to apply the set of relevant parameters for the representation of the universe through recording sessions
- a definition of the size of the resource and of each session

⁷ The monologue context can fit exactly with the definition of “only one intervenient” only in the formal use of language. In this case the social rules governing the interaction among participants may ensure the execution of the communication task by a sole speaker. On the contrary, in the everyday informal use of language, although only one speaker performs the main communication task, other participants may interact in the communication event with low informative contribution.

Given the variation parameters in I.1.1 the sampling strategy adopted for the representation of speech variability in the C-ORAL-BRASIL corpus is a function of the following principles:

- Definition of the size of the resource. Due to comparability constraint with the C-ORAL-ROM resource, a content of around 400.000 words (200,000 words for each category, formal and informal) was fixed. This size represents an improvement of more or less 35% on the C-ORAL-ROM size;
- Sampling of the universe with reference to *context variation* and to language *register variation*, leaving random speaker's variation⁸;
- Distinction between formal speech (50%) and informal speech (50%), thus ensuring a sufficient representation of dialogical Informal Speech (which is the resource with higher added value);
- The selection of the same criteria of C-ORAL-ROM for sampling the *formal* and *informal* part of the corpus;
- The definition of the text weight in terms of units of information (words);
- The definition of a text weight that ensures both the possible appreciation of macro-textual properties and sufficient representation of the universe in a 400.000-word corpus;
- The representation of a variety of possible recording situations within the range of perception and intelligibility of the human ear.

In a collection of a limited size, strong diatopical limits must be established. As C-ORAL-ROM does not represent in a systematic way diatopical phonetic variations due to the geographical origin of the speakers⁹, neither does C-ORAL-BRASIL. Assuming the relatively small size of the resource, the corpus design strategy concentrates on those variation parameters that, in principle, are more relevant to the documentation of speech variation and try to maximize the significance of the sampling for what regards the probability of occurrence of different types of speech acts and syntactic constructions. To this end, the greatest available different types of context of use and possible language tasks are represented, the more typical speech acts and modes of language construction in those contexts are represented. The diatopical documentation concentrates on the region of Minas Gerais, and especially on the metropolitan area of its capital, Belo Horizonte. This means that a large part, much more than 50%, of the speakers are from Minas Gerais. The statistics about speakers' origin are in DVD/Appendix/Measurements and in DVD/Book (chapter 2).

The use of the *formal / informal* parameter in the corpus design scheme allows the restriction of the number of significant parameters for what regards context variation. More specifically, while it can be assumed that in western societies the formal use of language is applied in a closed series of typical domains, the same does not hold for the informal use of language. The list of possible domains of use for informal language is by definition open, and no domain can in principle be considered more typical than others.

Under this assumption, the identification of the main domains of use of formal language maximizes the probability of representing the significant variations in this language variety, and is therefore the

⁸ The recording as part of the meta-data: a) Speaker characteristics; (gender, age, geographical region, education and occupation); b) acoustic quality of the text.

⁹ Corpora are mainly collected in Continental Portugal, Central Castile Spain, Southern France, Western Tuscany and are intended to represent a possible accepted standard, rather than all the varieties of pronunciation, which need collections of interlinguistic corpora with wide diatopical variability. This limitation is quite severe for Italian, where local varieties may strongly diverge from the standard (De Mauro et al., 1993). See in the Appendix. In the Book (chapter 2) the statistics on the geographical origin of speakers and detailed excel tables are presented in the Appendix.

best strategy. On the contrary, if significant variations of informal spontaneous speech are to be considered, the same strategy will cause a reduction of their probability of occurrence. For the documentation of the informal part, the variations in social context of use and in dialogue structure are the parameters systematically adopted, while the choice of the specific semantic domain of use is left random.

Also the strategy regarding the text weight varies its significance considering the Formal and in the Informal use of language. The formal use of language features in general long textual structures, while in the informal the length of syntactic constructions is limited. Therefore in order to ensure the probability of occurrence of typical structures, the text length for the Formal sampling must be significantly longer.

The above variation parameters and sampling strategy are projected in the corpus design matrix presented in the following paragraph.

1.4 Comparability

By definition, spontaneous speech comparability cannot be obtained through the use of parallel corpora. The resource of the C-ORAL-BRASIL corpus is comparable with the resources of C-ORAL-ROM as far as it satisfies the conditions on the corpus design stated in the following matrix, which reflects the variation parameters defined in 1.1

Section MANDATORY	Context MANDATORY	Domain MANDATORY	Number of words MANDATORY
INFORMAL [-Partially scripted]*			At least 150,000
	Family /Private context [-Public; -Partially scripted]		(124,500)
	Public context [+Public; -Partially scripted]		(25,500)
		Monologues	48,000
		Dialogues/Conversation	102,000
Section MANDATORY	Context MANDATORY	Domain MANDATORY	Number of words

C-ORAL-BRASIL number of words is higher than C-ORAL-ROM number of words for all the domains of both the contexts of the informal section; the percentage of public context is also a little higher in C-ORAL-BRASIL (23,43%) than in C-ORAL-ROM ($\pm 20\%$):

C-ORAL-BRASIL words distribution

Section	Context	Domain	Number of words
INFORMAL			208.130
	Family /Private context		(159,364)
		Monologues	52.116
		Dialogues	55.361
		Conversation	51.887
	Public context		(48,766)
		Monologues	16.222

		Dialogues	17.997
		Conversation	14.547

TEXT LENGTH REQUIREMENTS

In the informal section there are:

Normal texts: 110 texts of around 1500 words each (roughly from 1200 to 1800 words);

Short texts: 20 texts of less than 1100 words (the smallest one is of 167 words), but textually autonomous;

Long texts: 10 texts of more than 2000 words (the biggest one is of 4834 words).

Most of the length variation is in monologues.

For a corpus of the dimensions of C-ORAL-BRASIL the samples cannot be too big, in order to represent a sufficient variation. At the same time they cannot be too small, in order to guarantee a real textual status. Like in C-ORAL-ROM we recognize in around 1500 words a textual dimension that satisfies both the necessity of having a sufficient variation without losing the textual dimension. The few bigger texts, and specially the biggest one, represent some special situations that are too difficult but very important to be documented. The smaller texts represent, usually, textual types that do not reach the desired medium number of words: jokes, telling some specific episode of life, recipes are usually much smaller than 1500 words. The important thing is that all the texts are textually autonomous.

The term “word” refers to all graphic elements in the text surrounded by two spaces corresponding to the orthographic transcription of speech. All signs in the textual files corresponding to metadata, dialogue representation format and tagging elements are not counted as words.

Tolerance

The minimum requirements in the corpus design matrix concerning *section*, *context*, *domain structure* and *text length* are mandatory. The minimum target number of words requested for each field in the matrix has been approximated in each language collection.

In C-ORAL-BRASIL a big effort has been made to ensure the greatest variety of pragmatic situations. This means that the corpus was built by samples of different situations, both for dialogues/conversations and monologues. Especially for dialogues and conversations, the speakers were performing some activity, which means that their interaction was finalized to this activity. The participants of the interaction have been recorded during an activity and many times it was a moving activity, either on foot or by car. Movement (and different kinds of movements) is a very important feature in order to capture diaphasic variation. The number of texts in which two or more speakers are just engaged in a chat is reduced.

An important concept to be discussed is the difference between family/private and public context. The definition of family/private context is related to the fact that the speaker is acting in their familiar or private environment. First of all, family and private contexts are not exactly the same, but we assumed that in present western cultures there are few linguistic differences between the two contexts and they can be considered together, without looking for a specific balance between them. Other difficulties we had in defining the difference between private/ familiar on one hand and public on the other hand follow. For the Italian corpus, a speaker was considered as acting in a public context if his speech was exposed to be heard by unknown people and he/she was aware of it. In the BP corpus a speaker was considered as acting in a public context if he/she was acting not as a specific individual in a private role but in a social role. For instance, the speaker XXX acts as XXX with his friends or with his parents, but acts as a citizen toward a representative of the public administration, as a client toward the seller in a shop, as a professor if he discusses the thesis with

his student, even if other people do not take part in these interactions. Spanish, EP and French corpora seem to consider public context in even different ways.

The corpus design and the sampling criteria of C-ORAL-BRASIL ensure:

- The production of a corpus which offers, in principle, a representation of the wide variety of syntactic and prosodic features of spontaneous speech;
- The production of a resource comparable with the multilingual corpora (C-ORAL-ROM) as a basis for the induction of linguistic generalizations regarding spontaneous speech in the four Romance languages.

2. Filenames conventions

As for C-ORAL-ROM, filenames of the C-ORAL-BRASIL resource bear information of three types, in order:

- a) The represented language
- b) The text type; that is the field and sub-field to which each text belongs in the corpus structure
- c) The serial number identifying each text in its sub-field

The following are the conventions adopted in each C-ORAL-ROM and C-ORAL-BRASIL language collection:

1) Language

Country code: “f” (French), “i” (Italian), “p” (European Portuguese), “e” (Spanish), “b” (Brazilian Portuguese)

2) Text type

Informal section:

context: “fam” (family_private), “pub” (public)

domain: “mn” (monologues), “dl” (dialogues), “cv” (conversations),

3) Text identification number:

Two serial numbers identifying the text progressive in the sub-field;

e.g.:

bfamdl01 (Brazilian_Portuguese, family_private, dialogues, 01)

bfammn02 (Brazilian_Portuguese, family_private, monologues, 02)

bpubcv03 (Brazilian_Portuguese, public, conversation, 03)

3. Annotation information

For each recorded session, C-ORAL-BRASIL (like C-ORAL-ROM) provides the following set of annotations:

1. Metadata
2. Transcription and Dialogue Representation
3. Prosodic tagging of terminal and non-terminal breaks
4. Alignment
5. Part of speech (PoS) and lemma tagging of each transcribed form

3.1 Meta-Data

For each session, an ordered set of metadata is recorded as independent files in CHAT format.

The following are the rules for marking the metadata of C-ORAL-ROM in CHAT format, from which the annotations in other formats derive:

- Each metadata type is introduced by “@” immediately followed by a label, followed by “:” and an empty space.
- Metadata are listed in a closed set of types regarding
 - the session;
 - its size;
 - the speakers;
 - the acoustic quality;
 - the source;
 - the person who can provide information on the session.

Metadata fields or sub-fields that cannot be filled for a lack of information are filled with an 'x' (capital or small cap).

Label Type	Description
@Title:	A few words that help to recognize the text
@File:	Filename without extension (The name of audio file and the text file differ only in the extension).
@Participants:	Three capital letters identifying each speaker, followed by the corresponding proper name (first name), plus a sub-field with an ordered set of information on the speaker.
@Date:	Date of the recording: Day/month/year, separated by slashes; e.g. 20/06/2001 In a few files it was recovered only the month and year
@Place:	Name of the specific place and of the city where the recording session takes place
@Situation:	Ordered set of information: genre and role of the participants in the situation, environment, main actions performed, recording conditions; according with the rules specified below; e.g. <i>gossip between friends at home during dinner, not hidden, researcher participant</i>
@Topic:	The main argument dealt with in the speech event (max 50 characters); e.g. <i>traffic problems</i>
@Source:	Name of the collection leading to a copyright holder: C-ORAL-BRASIL
@Class:	The set of fields on the text class in accordance with the C-ORAL-ROM corpus structure (<i>separated by commas</i>)
@Length:	Length of the transcribed audio file in minutes(') and seconds (") e.g.: 12' 15"
@Words:	Number of words in the text file
@Acoustic_quality:	The acoustic quality of the recording. In accordance with specific criteria (A AB B BC or C)
@Transcriber:	Name of the person that selected and transcribed the text
@Reviser	Names of the revisers
@Comments:	Transcriber's or revisers comments on the text

- Each metadata type is filled with PCdata (closed or open vocabulary) ending with “enter” and can be specified in accordance with the rules.

3.1.2 Rules for the *Participant* field

Ordered set of sub-fields for each Participant (in parentheses, separated by a comma and an empty space).

Type	Description	Vocabulary	O(ptional)/ M(andatory)
Sex	Sex of the speaker	Closed vocabulary. One of the following conventional vocabularies according to the description in brackets: Man (Male) Woman (Female) X (unknown)	M
Age	Age of the speaker	Close vocabulary. One of the following conventional capital letters for each range of age between brackets : A (18-25) B (26-40) C (41-60) D (over 60) M (minor) X (unknown)	M
Education	The level of education according to the schooling degree	Closed vocabulary. One number according to the degree between brackets: 1 (primary school not completed or illiteracy); 2 (up to graduated that does not or did not use the graduation for her/his job) 3 (graduated who uses the graduation for her/his job or higher) X (unknown)	M
Profession:	Profession of the speaker at the moment of the recording	Open vocabulary. Name of the profession (e.g. professor; secretary; undergraduate student; etc.) or X (unknown);	M
Role	Role in the recorded event (even if it is equal to the profession)	Open vocabulary. Name of the role in the interaction (e.g. interviewer; interviewee; participant; intervenient) and type of relationship between participants; (e.g. father; professor)	M and O
Geographical origin/linguistic influence	Name of the city and state of speaker's origin	Open vocabulary (e.g. Belo Horizonte/MG; São Miguel d'Oeste/SC) or X (unknown) and information about other relevant linguist influences (e.g. lives in Belo Horizonte for 10 years)	M and O

3.1.3 Rules for the *Class* field

Informal

Type:

family/private

public

Sub-type:

Monologue

Dialogue

Conversation

3.1.4 Rules for the *Situation* field

FREE DESCRIPTION

The situation field is a set of information reported in discursive manner that helps to identify context in which the language event take place. The following are the guidelines used in C-ORAL-ROM and in C-ORAL-BRASIL to define the set of possible relevant information for the situation field*.

Type	Description	Vocabulary
Genre	Information that helps to define the genre of the linguistic event	Open vocabulary; e.g. (gossip; chat; quarrel; discussion; narration; claim, etc).
Reciprocal role	The reciprocal role of the participants	Open vocabulary (e.g. friends, colleagues, relatives)
Action	Main action performed during the speech event (if any)	Open vocabulary (e.g. while cooking, while playing snooker)
Recording conditions	Status of the recording with respect to the “Observer paradox” in spontaneous speech resources	Closed vocabulary. Two information aspects: 1) microphone not hidden 2) a choice of one alternative among researcher participant vs. researcher observant vs. researcher not present.

3.1.5 Rules for marking *Acoustic quality*

Texts in the collections are MANDATORILY labeled with respect to the acoustic quality of the sound source*. All the recordings are digital.

Properties	Label
Very high quality. Very good microphone response. Almost the entire recording is good for almost any kind of phonetic analysis. Almost no overlapping. Almost no background noise. F0 computing possible in (almost) the entire file.	A
High quality. Very good microphone response. The greatest part of the recording is good for almost any kind of phonetic analysis. Few overlapping. Little background noise. F0 computing possible in (almost) the entire file.	AB
Medium quality. Good or medium microphone response. Many parts of the recording are useful for phonetic analysis. F0 computing possible in the greatest part of the file.	B

Not too many overlaps and not too much background noise.	
Low quality. Medium microphone response. F0 computing possible in at least 60% of the recording. When it is not possible, the recording is still clear for the listener.	BC
Low quality. Medium or low microphone response. The F0 computing possible in at least 60% of the recording. Some parts of the recordings are not clear for the listener.	C

3.1.6 Quality assurance on metadata format

The format correctness of metadata has been double checked manually.

3.1.7 Statistics from C-ORAL-BRASIL metadata

A set of statistics on the C-ORAL-BRASIL metadata are reported in excel files in the DVD and in the book (chapter 2). These are the main figures regarding Speakers metadata and sessions metadata.

3.1.7.1 Number of speakers

Given that each database is substantially anonymous, the number of speakers is estimated on the metadata set. More specifically, a speaker is identified in each collection by the identity of short name, together with the long name, sex, age, and geographical origin, when these fields are filled with positive data¹⁰. It is possible that the same speaker is identified in different texts by different short names (three capital letters) and different speakers can be identified by the same short name. However, this does not create problems in the identification by the reader (the other metadata are always sufficient to disambiguate and there are no possible confusion in the same text) and to calculate the number of speaker: as a matter of fact, each speaker (independently of his short name) has an identification number. It cannot happen that the same identification number is given to different speakers or that the same speaker has two different identification numbers.

The number of speakers is given below, for Brazilian Portuguese of C-ORAL-BRASIL and for the four languages of C-ORAL-ROM. Notice that C-ORAL-ROM features both the informal and formal corpora, while C-ORAL-BRASIL so far shows only the informal part.¹¹

Brazilian Portuguese	ITALIAN	FRENCH	European Portuguese	SPANISH
Informal Speakers: 362	Total Speakers: 451 Informal: 209	Total Speakers: 305 Informal: 164	Total Speakers: 261 Informal: 106	Total Speakers : 410 Informal: 164

3.1.7.2 Distribution of speakers per geographical origin

The following is the distribution of speakers per geographical origin in C-ORAL-BRASIL informal and in C-ORAL-ROM, both informal and formal, according to the metadata. The high number of unknown speakers in C-ORAL-ROM is mainly due to the media collections where this information is not available (see distribution of absent speakers' metadata below). The high number of unknown

¹⁰ This restriction is intended not to overestimate the number of different speaker, given that some information about the same speaker might be not available to every transcriber.

¹¹ The number of different speakers in human-machine interaction cannot be computed.

speakers in C-ORAL-BRASIL is a consequence of the high diaphasic variation: in many different situations unforeseen people show up. Their weight in terms of words is very low: 2570 words, that is only 1,23% of the entire corpus.

It is noticeable that C-ORAL-BRASIL is also very homogeneous from the diatopic point of view. The corpus reflects the Mineiro variety and specially the variety of the metropolitan region of Belo Horizonte.

Brazilian Portuguese (informal)	speakers
Belo Horizonte	138
Other cities of Minas Gerais state	89
Other Brazilian states	19
Other countries	2
Unknown	114
Total	362

FRENCH	speakers
Provence and South of France	103
Poitiers and West France	28
Paris	26
Other regions	17
Centre of France	12
Other countries	5
Other Francophone countries	2
Unknown	112
Total	305

ITALIAN	speakers
Tuscany	188
South and Isles	38
Other countries	20
Central Italy	16
North - Various Regions	14
Unknown	175
Total	451

European Portuguese	speakers
Lisbon and Center Portugal	77
North Portugal	19
South Portugal	16
Açores and Madeira	10
Overseas	8
Other Regions	7
Other countries	5
Unknown	119
Total	261

SPANISH	speakers
----------------	-----------------

Madrid and Castillia	188
South America	19
Andalusia	11
Extremadura	11
Others Regions	11
Catalunia	7
Other countries	2
Unknown	161
Total	410

3.1.7.3 Completeness of speakers features in the metadata records

The following set of statistics is given for what regards the completeness of the information for the speakers reported in the metadata records of each session. Statistics refers only to the main features, i.e. *sex*, *age*, *education*, *geographical origin*, that may be significant for the exploitation of the resource¹².

The number of recordings where the information is complete are identified and for each language corpus and in the main field of the corpus design.¹³ Moreover the percentage of recordings in which each of the main features is unknown is also reported for each language corpus and for each main field in the corpus design.¹⁴

Brazilian Portuguese

INFORMAL

RECORD	362
SEX-AGE-ORIGIN-EDUCATION	68,23%
NO SEX	0,28%
NO AGE	29,55%
NO ORIGIN	31,49%
NO EDUCATION	30,38%

French

TOTAL INFORMAL NATURAL CONTEXT MEDIA TELEPHONE PRIVATE

RECORD	456	243	64	91	58
SEX-AGE-ORIGIN-EDUCATION	57,89%	75,72%	48,44%	9,89%	68,97%
NO SEX	0,00%	0,00%	0,00%	0,00%	0,00%
NO AGE	24,56%	13,17%	37,50%	42,86%	29,31%
NO ORIGIN	36,40%	16,87%	48,44%	84,62%	29,31%
NO EDUCATION	24,78%	10,29%	18,75%	63,74%	31,03%

Italian

TOTAL INFORMAL NATURAL CONTEXT MEDIA TELEPHONE PRIVATE

RECORD	596	273	70	214	39
SEX-AGE-ORIGIN-EDUCATION	46,14%	75,09%	40,00%	5,14%	79,49%
NO SEX	0,00%	0,00%	0,00%	0,00%	0,00%

¹² Profession and role fields are considered additional information and are not object of this statistics evaluation.

¹³ In C-ORAL-ROM, the number of record is superior to the number of speakers, given that the same speaker may appear in different sessions with more or less metadata information available. This does not happen in C-ORAL-BRASIL.

¹⁴ This information is never available for Speakers of Human Machine interactions and for speakers of mixed turns that therefore have not been counted.

NO AGE	24,50%	12,09%	30,00%	42,99%	0,00%
NO ORIGIN	37,58%	3,66%	41,43%	84,58%	10,26%
NO EDUCATION	36,24%	20,51%	20,00%	66,36%	10,26%

Portuguese

	TOTAL	INFORMAL	NATURAL	CONTEXT	MEDIA	TELEPHONE	PRIVATE
RECORD	437	188	103	109			37
SEX-AGE-ORIGIN-EDUCATION	54,92%	97,34%	16,50%	2,75%			100,00%
NO SEX	0,00%	0,00%	0,00%	0,00%			0,00%
NO AGE	40,96%	0,00%	75,73%	92,66%			0,00%
NO ORIGIN	43,48%	2,66%	76,70%	97,25%			0,00%
NO EDUCATION	34,78%	0,00%	55,34%	87,16%			0,00%

Spanish

	TOTAL	INFORMAL	NATURAL	CONTEXT	MEDIA	TELEPHONE	PRIVATE
RECORD	553	204	58	269			22
SEX-AGE-ORIGIN-EDUCATION	51,36%	100,00%	65,52%	7,43%			100,00%
NO SEX	0,00%	0,00%	0,00%	0,00%			0,00%
NO AGE	36,71%	0,00%	5,17%	74,35%			0,00%
NO ORIGIN	44,85%	0,00%	34,48%	84,76%			0,00%
NO EDUCATION	0,18%	0,00%	0,00%	0,37%			0,00%

Statistics shows that only “sex” is always filled (just one case in BP is not filled: this speaker just pronounces an exclamation and the voice does not render the sex clearly). However the relevant information regarding the speakers is complete in all or most of the Informal and Telephone sub-corpora that are the more significant contexts for this type of information. In BP the percentage of speakers lacking some or all relevant information is higher because, as it was said, the number of unknown people that appear in the recording is very high (114, for just the informal session) but reflect a very small percentage of the pronounced words (1,23%). For C-ORAL-ROM, it must be considered that this kind of information is usually not available for media sessions and only occasionally available in Formal in natural_context sessions.

3.1.7.4 Completeness of session metadata

All session metadata are filled with real information. No section is filled with empty information.

3.2. Transcription and Dialogue representation

The C-ORAL-BRASIL dialogue representation is defined as an implementation of the CHAT architecture (MacWhinney, 1994) (<http://chilides.psy.cmu.edu/manuals/CHAT.pdf>) and has the following structure:

Text lines: orthographic transcription of the speech information divided:

- Vertically*, in dialogic turns (introduced by a speaker label)
- Horizontally*, by prosodic parsing and utterance limit, representing terminal and non terminal prosodic breaks of the speech continuum

3.2.1 Basic Concepts for dialogue representation

Concept	Definition
Dialogic Turn	Continuous set of speech events by only one speaker's voice . The dialogic turn changes if, and only if, a speech event by another speaker occurs.
Session	Set of dialogic turns corresponding to one meta-data set.
Utterance	The minimal speech event by a single speaker such that it can be pragmatically interpreted as a <i>speech act</i> . ¹⁵
Word	A speech event perceived as a phonetic unit, such that it conveys a meaning.

3.2.2. Turn representation.

A dialogic turn by one speaker is expressed by “*” immediately followed by three capital letters identifying the speakers in the metadata, then followed by “:” and one space before the transcription of the speech event. Each dialogic turn ends with an “enter”.

Convention ¹⁶	Description
^*[A-ZÑ]{3}:\s{1}	Dialogic turn of a given speaker

3.2.3 Utterance representation.

Each utterance is represented by a series of transcribed words ending with the termination symbol “//” or other symbols having a termination value (see below).

E.g.:

*MOR: I'm going home // I'm tired //
 *PIL: bye bye //

A dialogic turn can also be filled with *non linguistic* or *paralinguistic* material according to the transcription convention (see below)¹⁷

*MAX: are you sure //
 *PIL: hhh

In C-ORAL-BRASIL it was decided that when some paralinguistic material plays a specific communicative and conventionalized role, even if it is in isolation, it is considered as an utterance and followed by double slash. If it does not happen, the transcriber, whenever possible, transcribes the *hhh* inside the closest utterance of the speaker.

3.2.4. Word representation.

Each word is transcribed as a continuous sequence of characters between two empty spaces, in accordance with the orthographic convention of each language.

3.2.5 Transcription

1. The transcription of a dialogic turn expresses, horizontally, the sequence of speech events (utterances) that occur in each dialogic turn; i.e., in principle, no “enter” can occur within an utterance. This may not be the case in *overlapping* and other *cross over dialogue* phenomena (See below).

¹⁵ See the operative definition of Speech Act in § 3.4.3 and references therein.

¹⁶ Expressed through regular expression notation.

¹⁷ In this case the dialogic turn is filled by a communication event instead of a speech event. The relation between the two concepts is left undefined in C-ORAL-BRASIL; this resource marks the main communication events in a dialogue, but is not specifically devoted to the study of such events, which must be considered in a multimodal framework.

2. The transcription follows the standard orthography of each language, and is integrated with special signs devoted to handling spoken language phenomena. No phonetic transcription is found. Orthographic choices made in the Brazilian Portuguese collection are reported in APPENDIX 3.

C-ORAL-BRASIL made a great effort to integrate the traditional orthography with a graphic system that could capture a high quantity of those speech phenomena that are candidate to grammaticalize or lexicalize.

Concept	Definition
Text	Each recorded session is an ordered collection of transcribed dialogic turns referring to a given metadata set.

3.2.6 Overlapping

The speech of one speaker in spontaneous dialogues is frequently overlapped by the speech of another speaker, who may insert his dialogic turn in the other speaker's turn. The overlapping therefore determines a relation of temporal equivalence between two or more speech portions in different dialogic turns. In C-ORAL-BRASIL, overlapping is represented through the conventions reported below.

The overlapped text in both dialogic turns is placed between brackets < >.

Overlapping is marked only when at least two words in two different turns are concerned; this means that the overlapping of syllables is left unmarked or reported to word boundaries.

When, due to the simultaneous occurrence of more than one dialogic turn, it is impossible to attribute the speech to a speaker, a fixed variable is used to mark a mixed-turn, e.g.:

*XYZ: <quem jogou isso> //¹⁸ [who threw this]

OVERLAPPING

Symbol	Description
< >	Brackets which mark the beginning and the end of the overlapped text of a given speaker
*XYZ:	Turn of overlapped speech by non-identified speakers

The speech software allows the alignment of the overlapped segments on independent layers of the different speakers.

3.2.7 Cross over dialogue convention

In spontaneous spoken language, the event of an intersection of dialogic turns by different speakers may occur; this means that a dialogic turn may arise *before* the end of the turn that immediately precedes it.

Therefore, in those cases, the representation of dialogue as a vertical ordered collection of dialogic turns may be maintained with difficulty. Because the C-ORAL-BRASIL (like the C-ORAL-ROM) format forces the representation of the sequence of turns in a temporal order, a *cross-over dialogue convention* has been adopted.

Cross-over dialogue convention

In the transcripts, a slash placed at the beginning of a turn, i.e. immediately after the turn label and before the transcribed text, is a convention expressing that the turn in focus is only virtual, while the linguistic information of the turn belongs to the preceding turn of the same speaker¹⁹

¹⁸ In these cases the transcription may not be reliable for perceptual reasons.

¹⁹ Each slash immediately following the speaker mark is not counted as a prosodic break.

Configuration of symbols	Operational definition
: /	Relation that converts the turn in which it appears into a linear sequence which includes the preceding turn of the same speaker

Three major cases of cross-over dialogue have been detected in C-ORAL-ROM:

- 1) overlapping;
- 2) interruption by the listener
- 3) complete intersection of turns

3.2.7.1 Overlapping and cross-over dialogue

Because overlapping is a relation between texts belonging to different turns, it affects the dialogue representation, which expresses the time dimension both in vertical and horizontal relations. In principle, the dialogue representation system requires the linguistic information which follows vertically in a subsequent turn to be also necessarily subsequent in time to the text reported in the previous turn. However, this cannot be the case in most overlapped sequences, where a dialogic turn continues despite the insertion of another turn in it that partially overlaps it, e.g.:

*ABA: Ela me falou [/] <que não vai> mais / ao show // porque não tá bem //
 *BAB: <vai> //

3.2.7.2 Intersection of turns

Some cases of complete intersection of dialogic turns may occur, without interruption or overlapping; e.g. in the following example, the listener, without really interrupting the speaker, starts a brief dialogic turn that goes on with his prosodic program.²⁰ Even if overlapping is absent, or reduced to a few milliseconds, the speaker may interrupt his utterance in connection with the intervention of the listener. For example, in the following situation, a speaker inserts himself in a dialogic turn, interrupting it, but the other speaker goes on with his turn despite the interruption:

*ABC: eu acho que / as coisas /
 *CBA: são boas //
 *ABC: / estão indo bem //

3.2.8 Transcription conventions for segmental features

C-ORAL-BRASIL has developed a set of non-orthographic criteria in order to capture the main phenomena of lexicalization and grammaticalization of Brazilian Portuguese and to analyze them with quantitative and statistic criteria (Mello & Raso 2009). The different decisions are discussed in chapter 5 of the Book. As orthographic standard we use the standard of Novo Dicionário Houaiss in the CD-ROM version (1990).

3.2.8.1. Other transcription conventions

3.2.8.1.1 Hesitations and interrupted words

Speaker's hesitations and voiced time taking are transcribed as *&he*, independently of the vocal quality.

3.2.8.1.2 Onomatopoeias

Onomatopoeias are transcribed tentatively, according to the pronunciation.

3.2.8.1.3 Interjections and exclamations

The interjections *ah*, *eh*, *ih*, *oh* e *uh* are orthographically transcribed.

²⁰ In this case the utterance-based alignment cannot be maintained.

Some interjections are used with high frequency to confirm, to deny or to ask for confirmation or repetition. To confirm: *hum hum*, *ham ham*, *hum e ham*. To deny: *uhn uhn e ahn ahn*. To ask for confirmation or repetition: *uhn e ahn*.

Religious exclamations are transcribed with capital initial. Some of them have reduction forms: *Nossa Senhora* can become *Nossa*; *No'*, *Nu'* e *Nusga*. *Virgem Maria* can become *Vixe'* e *Vix'*. *Ave Maria* can become *Aff'*.

3.2.8.1.4 Abbreviations and acronyms

Abbreviations and acronyms are transcribed orthographically and with lower case when they are pronounced as words. They are transcribed with lower case and as they are spelled if they are the spelling of letters (ex. *Ufeemegê* for UFMG).

If they are already considered as words, they are transcribed orthographically (ex. *radar*).

When they are followed by a number, the numbers are always transcribed after them as a different word, independently whether they are or are not part of the acronym (e.g.: *emepegue um*).

The form “OK” is transcribed *oquei*.

3.2.8.1.5 Numerals

If they are formed just by one word, they are transcribed orthographically. If they are formed by more than one word, they are transcribed separated by hyphen.

3.2.8.1.6 Foreign words and mispronunciations

When there is an orthographic tradition in Portuguese for a given foreign word, it is transcribed according with this tradition (e.g. *estresse*), unless it is clear that the speaker is deliberately using the original pronunciation. When a Portuguese orthographic tradition is absent, it is used the original orthography.

Names of foreign products (*YouTube*, *Google Video*) are transcribed with the original orthography, even if there is a Portuguese version, as in the case of *Google Video*, for which there is the Portuguese orthography *video*.

If the pronunciation of the foreign word is not correct (ex. *Big Brogher* instead of *Big Brother*), the real pronunciation is transcribed, and the mistake is reported in the comments of the metadata. As far as mistaken pronunciations are concerned in Portuguese or in whatever language, if the speaker uses an incorrect pronunciation and then corrects himself, the transcriptions report both the wrong and the correct pronunciations. If the speaker uses a wrong pronunciation and does not correct himself, just the wrong pronunciation is transcribed and the mistake is reported in the comments of the metadata.

In the case of the expression *et cetera*, it was decided it would be transcribed as one unique word: *etcetera*.

3.2.8.1.7 Forms that are transcribed following the standard orthography even if pronounced differently

The standard orthography is always used for:

1. Infinitives: even when the final *-r* is not pronounced.
2. Letters of the alphabet: *á*, *bê*, *cê*, *xis*, *jota* etc.
3. The forms *que*, *quê*, *por que*, *porque*, *porquê*.
4. Titles and proper nouns are always transcribed with capital initial: *ele apareceu no Programa do Jô*.
5. Citations: are transcribed between double quotation marks. Reading and singings are treated as citations.

3.2.8.2 Non-orthographic criteria

3.2.8.2.1 Apheresis

Some examples: *brigado*<*obrigado*; *güentar*<*agüentar*; *fessora*<*professora*; *bora*<*embora*. All the apheretic forms of the verb *estar* are transcribed as pronounced: *tô*, *tava*, *tá*, *tando*, *tar*, *tão* etc., even the non-standard form *teje*.

3.2.8.2.2 Verbal conjugation

First person plural of verb *ir* may be transcribed *vamos*, *vamo* and *vão*, depending on pronunciation. The infinitive *vir* is transcribed *vim* when is pronounced so.

For verb *ter*, it is frequent the pronunciation *tem* for *tenho*, either in the periphrasis *tenho que/tenho de*, or alone. The pronunciation *tem* is always maintained in transcriptions.

The apocopated forms for *pode*, like *po'* (as in *po' fazer*) are transcribed *po'*.

The form *deixa* of the verb *deixar* is transcribed orthographically when it is pronounced as *deixa* or *eixa*, but it is transcribed *xá* when there is apheresis of the first syllable.

The form *tó*, of the verb *tomar* in the imperative (equivalent to *toma/tome*), meaning transfer of possession is transcribed as pronounced.

The verb *olhar* in the indicative may be reduced in expressions like *olha lá>o' lá* or *a' lá*. These two reduced forms are preserved in transcriptions.

In all verbs, first person plural is transcribed respecting pronunciation. For example *empurramos* or *empurramo*, or *empurremo*. The same in specific forms of different persons like *seje* for *seja*, or in the radical of specific verbs like *envem* for *vem*.

The reduction of all verbal paradigms is always respected in transcriptions, like in *nós vai*, *eles foi*, *vocês faz*, *tu é* etc.

3.2.8.2.3 Plural marks

The pronunciation without plural marks is respected in transcription, like in: *os carro*, *os menino bonito* etc.

When invariable exclamations are pronounced with plural marks, these are transcribed, like in *ques menino bonito* (= que meninos bonitos), *ois menino* (saudação = oi meninos), *ôs menino* (chamamento = ô meninos).

3.2.8.2.4 Pronouns

The subject personal pronouns cliticization (or weakening) is an important grammaticalization phenomenon going on in Brazilian Portuguese. The transcription criteria foresee two series for second and third person singular and plural. Pronouns are transcribed as pronounced: the first series is *vocês/ocê*, *ele/ela*; *vocês/ocês*, *eles/elas*; the second one is *cê*; *e'/ea*; *cês*; *es/eas*.

Demonstratives have two series as well: the first one is *aquele*, *aqueles*, *aquela*, *aquelas*; the second one *aque'*, *aque*s, *aquea*, *aqueas*.

3.2.8.2.5 Articulated prepositions

Articulated preposition contractions are respected in transcriptions: *pra*, *pro*, *pras*, *pros*, *prum*, *pruma*, *pruns*, *prumas*; *no*, *na*, *nos*, *nas*, *num*, *numa*, *nuns*, *numas*; *do*, *da*, *dos*, *das*, *dum*, *duma*, *duns*, *dumas*; *co*, *ca*, *cos*, *cas*, *cum*, *cuma*, *cuns*, *cumas* etc. The criteria foresee also the following forms: *pa*, *pas* (*para a*, *para as*), *po*, *pos* (*para o*, *para os*), *pum*, *puns* (*para um*, *para uns*) e *puma*, *pumas* (*para uma*, *para umas*).

Preposition *em* is transcribed *ni* when pronounced so (*ni mim*, *ni entrevista* are equivalent to *em mim*, *em entrevista*).

The criteria also foresee the transcription of reduced preposition forms: *para* can be pronounced and transcribed *pr'* or *p'*; *em* can be transcribed *n'*; *de* can be transcribed *d'*; *com* can be transcribed *c'* (see 3.2.8.2.6).

3.2.8.2.6 Articulated prepositions and pronouns

The criteria foresee the following possibilities for prepositions *para* (*pra*, *pa*, *pr'*, *p'*), *em* (*ni*, *n'*), *de* (*d'*), *com* (*c'*) when meeting pronouns *e'*, *ea*, *es*, *eas*, *ocê*, *ocês*, *cê*, *cês*. In case of *des* (for *deles*), or *nes* (for *neles*), as PB has the orthographic form *deles* and *neles*, the convention is to transcribe *des* and *nes* in one word. As a consequence, the criteria foresee the forms *de'*, *dea*, *des*, *deas*; *ne'*, *nea*, *nes*, *neas*. All other cases are treated differently: *pr' es*, *c' es* etc. As we have the forms *p'* and *cê*, it is possible to transcribe *p' cê*, and the same for *c' ocê* and *c' cê* if this is the pronunciation.

The compositional criterion makes it not necessary to foresee *a priori* all the possible forms. So, it is possible to follow the same criterion for the junction of prepositions and demonstratives: as the standard orthography has the forms *daqueles* and *naqueles*, the contracted forms will be *daques* and *naques*, and there will be *daque'*, *naquea* etc. When the one Word form is not accepted in the traditional orthography, the compositional criterion gives the solution with apostrophe plus space: *c' aques*, *pr' aques* etc.

3.2.8.2.7 Negation

Frequently negation *não* is in clitic or weak form. We created a graphic form (*nũ*) to report the phenomenon.

For the sequence *nénão*, there is the conventionalized form *n' é não*.

3.2.8.2.8 Interrogative constructions, relative and pseudo-relative pronoun

In interrogative constructions like *que que*, *por que que*, *quando que*, *quanto que* etc., the verb *é* is not transcribed if it is not pronounced. If the pronunciation of the first item ends with an open *e*, the verb is taken as pronounced and the transcription is *que é que*.

A similar phenomenon happens in cleft constructions. A sequence like *Maria é que faz* can be pronounced *Maria que faz*, and in this case the verb *é* will not be transcribed.

Frequently, after a demonstrative or in some other cases, the relative *que* is not pronounced, like in *esse (que) cê tá experimentando*, *esse (que) tá na sua mão*. If nothing leads to the perception of the *que*, it is not transcribed.

The expression of scaring or surprising or other attitudes *que é isso* or *que isso* can be pronounced with or without the verb. If the *que* is pronounced with open *e* the verb is transcribed, if the *que* is pronounced with closed *e* the verb is not transcribed.

3.2.8.2.9 The formal forms Senhor and Senhora

The form *senhor* is transcribed as it is pronounced, according to the following possibilities: *senhor*, *sior*, *seu*, *sô*. The pronunciation *sinhô* is transcribed *senhor*. *Senhora* is transcribed *siora*, *sio'* or *sá*, according to the way it is pronounced.

3.2.8.2.10 Diminutives

The apocopated diminutives are reported in the non-orthographic transcription criteria. For example, *sozinho* is transcribed *sozim* when this is the pronunciation. Diminutives may present more than one form, like in *devagarinho* e *devagarzinho*.

3.2.8.2.11 The intensifier mó

The form *mó*, meaning *maior* or *muito*, is transcribed as it is pronounced: *mó*.

3.2.8.2.12 Rotacism

Rotacism is reported in transcription, like in *Craudia* for *Claudia*.

3.2.8.3. Non-understandable words

All words that are not properly understood are reported (and counted as word occurrences in a frequency list) as:

xxx

3.2.8.4 Paralinguistic elements

All paralinguistic elements (laughing, crying, etc) are not counted as a word occurrence in a frequency list and are indicated as:

hhh

3.2.8.5. Fragments

All incomplete words and/or phonetic fragments are immediately preceded by *&*, as in the following incomplete utterance:²¹

*ABC: meu &ir [my &br]

or in the following retracting:

*ABC: meu &ir [/2] meu primo não é daqui // [my &br [/2] my cousin is not from here]

or in the following lengthening of the programming time:²²

²¹ Incomplete words are never subject to rebuilding, except, of course, for what regards systematic phonetic phenomena (elision, breaking off of the last syllable etc.). Those phenomena are mirrored (or not) in the transcription, following the orthography of each language and the particular traditions in editing oral text. Such choices must be detailed in the notes to the corpus edition.

*ABC: acho que o nome dele / &eh / é João // [I think that his name / &eh / is João]

When there is a retracting phenomenon C-ORAL-BRASIL shows, inside the brackets and after the slash, a number that shows the amount of words virtually deleted by the speaker:

*ANE: <pra &ba> [/2] pra baixo //

3.2.8.6. Interjections

Interjections are not fragments; they are phonetic elements with dialogical function. Interjections are transcribed following the lexicographical tradition of each romance language. New interjections discovered in the corpus are transcribed tentatively and their presence is reported in a glossary in the appendix.

3.2.8.7. Non standard words

Non-standard words found in the corpus are transcribed tentatively and their presence is reported in a glossary added in the appendix.

3.2.8.8. Non-transcribed words

When a word must be cancelled for reasons concerning privacy or decency it is substituted by a variable, “yyy”, to be counted as a word²³:

*MOR: o yyy é um idiota //
[yyy is an idiot //]

3.2.8.9. Non-transcribed audio signal

When, for whatever reason, part of the audio cannot be transcribed, a single variable “yyyy” is inserted in the transcripts, not depending on the length of the signal. The variable is aligned, but will not be counted as a word:

Symbol	Description
&	Mark for speech fragments
hhh	Paralinguistic or non linguistic element
xxx	Non-understandable word
yyy	Non-transcribed word
yyyy	Non-transcribed audio signal

3.2.9 Quality assurance on format and orthographic transcription

The format correctness of both metadata and transcripts has been checked manually at least three times.

The project ensures maximum accuracy in the transcripts that have been compiled by graduate students or by undergraduate selected for a specific fellowship. The original transcripts have been revised at least four times. The last two revisions were made by Ph.D. students transcribers. Orthographic conventions and non-standard forms have been registered in Appendix 3 of these specifications. The validation process for the transcriptions has been made by three transcribers. Results of the validation are reported in Appendix 5.

²² Note that the lengthening of syllables, which is quite a common and perceptively relevant phenomenon in spoken language, is not marked in this system. However, following the philosophy of marking prosodic breaks, the system automatic assumes the generalization that lengthening necessarily causes a prosodic break.

²³ When a word is not transcribed in the text it is substituted with some beep of a similar length in the acoustic signal.

3.2.10 Transcription validation

There were two validation process: one before (initial validation) and the other after (final validation) the last transcription revision. The text correctness and accuracy evaluation were carried by expert transcribers who checked on the aligned text to audio files through the software WinPitch²⁴.

The transcriptions were checked for the following problems:

- General mistakes:
 - Orthography mistakes; mistyping or transcription of form different from what is found in the audio signal;
 - Word insertion (inexistent word in the audio transcribed);
 - Word deletion (word present in the audio not transcribed);
- Incorrect application of transcription criteria:
 - The word should have been transcribed following standard orthography but was transcribed following non-orthographic criteria;
 - The word should have been transcribed following non-orthographic criteria but was transcribed following standard orthography;
 - Incorrect application of non-orthographic criteria resulting in non-predicted form in both orthographic and non-orthographic criteria.

The checked files did not have speech overlap. The examined samples were extracted at random, without data replacing.

3.2.10.1 Initial validation

The initial validation was carried in two samples as described in the table below. Sample number 1 was checked for all types of mistakes, both general and incorrect criteria application. In sample 1B only mistakes related to inadequate application of transcription criteria were observed.

Initial validation samples table

Sample	Size	Check
1A	7484 words (5% utterances from 89 texts)	All types of mistakes
1B	8949 words (10% utterances from 50 texts)	19 non-orthographic criteria

Results:

Sample 1A validation results – all types of mistakes search

Mistake type	Initial validation - sample 1A	
	mistakes/words	%
All the mistakes (1 + 2)	140/7484	1.87%
1. General mistakes (a +b + c)	104/6319	1.65%
a. Spelling/ inconsistency spelling-audio	45/6319	0.71%
b. Word insertion	19/6319	0.30%
c. Word deletion	40/6319	0.63%
2. Mistake in the application non-orthographic criteria	37/1165	3.18%

²⁴ Martin, P. (2011). WinpitchW7. Pitch Instruments. Retrieved from <<http://www.winpitch.com/>>.

The 95% confidence interval for transcription accuracy after the initial C-ORAL-BRASIL validation is placed within 0.978 e 0.984.

In the verification of sample 1B there were 1,308 words to which non-orthographic criteria were applicable and there were 8 mistakes found, whereby there was a 0.61% mistake rate.

3.2.10.2 Final validation

The final validation was undertaken in four samples, as shown below. In sample 2A, all types of mistakes were checked, both general mistakes and incorrect transcription criteria application mistakes. In sample 2B only inadequate transcription criteria application were observed. In sample 2C only mistakes related to transcription criteria affecting less than 90 tokens were observed. In sample 2D only mistakes related to a single transcription criterion underrepresented in other samples were examined.

Final validation sample table

2A	8243 words (5% utterances from 89 texts)	final	All types of mistakes
2B	8877 words (10% utterances from 50 texts)	final	19 non-orthographic criteria
2C	11805 words (5% utterances from 139 texts)	final	3 non-orthographic criteria
2D	11215 words (5% utterances from 139 texts)	final	1 non-orthographic criterion

Results:

Sample 2A validation results. Search for all types of mistakes

Mistake type	Final validation – sample 2A	
	mistakes/words	%
All mistakes (1 + 2)	67/8243	0.81%
1. General mistakes (a + b + c)	55/6124	0.90%
a. Spelling mistake / inconsistency spelling-audio	23/6124	0.38%
b. Word insertion	19/6124	0.31%
c. Word deletion	13/6124	0.21%
2. Non-orthographic criteria application mistake	12/2119	0.57%

The 95% confidence interval for the transcription accuracy after the last C-ORAL-BRASIL validation is placed within 0.989 e 0.993.

In sample 2B there were 693 words to which non-orthographic transcription criteria were applicable and there were 3 mistakes found, whereby there was a 0.43% mistake rate.

Sample 2C analysis only focused on three non-orthographic criteria (apheresis, preposition reduction and first person plural form). 66 words to which the criteria were applicable underwent examination (40 apheretic forms, 14 first person plural forms and 12 prepositions). Only one transcription mistake was found – it was a preposition transcribed in reduced form instead of its full form in orthographic spelling. The mistake rate in sample 2C was 1,5 % (1 mistake in 66 words).

Sample 2D analysis focused on preposition spelling. 16 reduced prepositions (c', p', pr', d', n') and 23 fully orthographically transcribed ones were examined. No mistakes were found in this sample.

3.3. Prosodic annotation scheme

3.3.1 Principles

C-ORAL-BRASIL prosodic tagging is informed by the following series of principles:

- The prosodic tagging specifies each perceptively relevant prosodic break in the speech continuum.
- All positions between two words are considered possible positions to be fitted with a prosodic tag. No within-word prosodic breaks are marked in C-ORAL-ROM.
- Prosodic breaks are distinguished in accordance with two main qualities: *terminal* vs. *non-terminal*.
- Each between-words position necessarily has one of the following values with respect to the prosodic tagging of the resource:
 - no break
 - terminal break
 - non-terminal break
- Prosodic breaks are always tagged and reported according to perceptual judgments of the transcribers, within the process of corpus revision and transcription accuracy.
- Prosodic tagging is part of the transcription and is reported within the text lines.
- The criterion for the segmentation of the speech flow into utterances is prosodic. Each prosodic break qualified as terminal defines the utterance limits in the speech flow

3.3.2 Concepts

Concept	Definition
Prosodic break	Perceptively relevant prosodic variation in the speech continuum such that it causes the parsing of the continuum into discrete prosodic units.
Terminal prosodic breaks	Given a sequence of one or more prosodic units, a prosodic break is known as terminal if a competent speaker assigns the quality of <i>concluding such sequence</i> to it.
Non-terminal prosodic breaks	Given a sequence of one or more prosodic units, a prosodic break is known as non-terminal if a competent speaker assigns the quality of being <i>non-conclusive</i> to it.
Prosodic pattern (Utterance)	Each sequence of prosodic units (≥ 1) ending with a terminal prosodic break ²⁵

3.3.3 Theoretical background

Studies have shown that perception is highly sensitive to voluntary F0 variation ('t Hart et al., 1990). In accordance with this theoretical framework, the melodic pattern which scans the speech flow is an object of perception. Each tone unit of a prosodic pattern corresponds to a perceptually relevant pitch movement. A prosodic pattern may be simple (composed of a single tone unit) or complex (in which case it is made up of two or more tone units melodically linked together).

From another point of view, according to the speech act theory tradition, every utterance in spoken language is the voluntary accomplishment of a speech act (Austin, 1962).

The background theory of the transcription format (Cresti, 1994, 2000) links the two properties: voluntary F0 variations do not simply scan the utterance, but rather express the information

²⁵ Heuristic definition of utterance in C-ORAL-ROM and C-ORAL-BRASIL.

necessary to the accomplishment of speech acts. For this reason, the selection of textual units corresponding to an utterance can be based on prosodic properties.

More specifically, it is possible to identify an utterance each time prosody enables the perception of the completion of a speech act; i.e. intonation permits the pragmatic interpretation of the text (*Illocutionary criterion* Cresti, 1994, 2000).

In the transcription format, the identification of utterances in the sound continuum is linked to the detection of perceptively relevant F0 movements with a terminal value. It is assumed that there is no such thing as an utterance without a profile of *terminal intonation* (Karcevsky, 1931; Crystal, 1975). Non-terminal tone units correspond to the scanning of an utterance by means of a complex pattern.

In other words, the systematic correlation between terminal breaks and utterance limits is the heuristic method for speech segmentation in *utterances*, that is, the segmentation of the linguistic information in the resource with respect to the specific unit of analysis of spontaneous speech (See. Miller & Weinert, 1998; Quirk et alii, 1985; Biber et alii, 1999; Cresti, 2000).

3.3.4 Conventions for prosodic tagging in the transcripts: types of prosodic breaks

To discriminate between terminal and non-terminal breaks is mandatory in C-ORAL-BRASIL and in all the C-ORAL-ROM transcripts. However, the format allows the prosodic tagging to be displayed at two hierarchical levels, with greater or lesser attention to:

- the annotation of the types of terminal breaks;
- fragmentation phenomena in the speech performance.

3.3.4.1 Terminal breaks (utterance limit)

A signal is inserted in the transcription each time a prosodic break is perceived as terminal by a competent speaker. Each terminal break indicates the prosodic completion of the utterance.

C-ORAL-BRASIL, differently from C-ORAL-ROM, chose to also mark interrogative, exclamative and intentionally suspended utterances with the generic terminal break “//”, with no supplementary specification. The reason for this is that the analysis of the type of illocution is a different task. The terminal break just marks the conclusion of the utterance that, of course, implies the performance of an illocution²⁶.

Value	Description	Symbol
<u>all possible illocutionary values</u>	<i>Concluding prosodic break</i>	//

3.3.4.2 Non-terminal breaks

The symbol “/” (single slash) is inserted in the transcription to mark the internal prosodic parsing of a textual string which ends with a terminal break; it is inserted in the position where a prosodic break, which is not perceived as terminal, is detected in the speech flow by a competent speaker.

Value	Description	Symbol
Non-terminal	Non-conclusive prosodic break	/

²⁶ In C-ORAL-ROM were distinguished different types of terminal breaks (“//”, “...” and “?”) for concluded utterances.

3.3.5 Fragmentation phenomena

The annotation scheme embodies the generalization that a prosodic break always occurs when a fragmentation of the linguistic information arises in the speech performance. That is, in spontaneous speech, a break of the prosodic unit in which the fragmentation arises.

When a prosodic break occurs in connection with a fragmentation phenomenon, at the richer level of transcription, the prosodic tagging is specified in accordance with the complete set of the following alternatives:

Symbol	Description	Type of break
+	<i>Concluding prosodic break such that the utterance is interrupted by the listener or by the speaker himself</i>	Terminal
[/n]	<i>Non-conclusive prosodic break caused by retracting. The n indicates the number of words cancelled by the speaker</i>	Non-terminal

3.3.5.1 Interruptions

The interruption (non-completion) of an utterance may be due to any reason: a change of the linguistic programming by the speaker, an interruption caused by the listener or by other events in the environment. Interruptions may be accompanied by word fragmentation (interruption before the end of the last word of the utterance) or, as is more frequently the case, may not feature any word fragmentation.

*interruption mark*²⁷: +

The interruption mark is counted as a type of terminal break. The sign is inserted in the transcription in the position where the utterance is interrupted because of an interruption made by the listener, or because of a change in programming by the speaker (Examples in Appendix 1)

3.3.5.2 Retracting and/or restart and/or false start(s)

The retracting phenomenon (or false start) is the most frequent fragmentation phenomenon in spontaneous speech. The speaker hesitates while trying to find the best way to express himself and retracts his speech before choosing between two alternatives. This phenomenon is, generally, clearly distinguishable from interruptions or changes in programming, reported above, which do not feature speakers' hesitations. Contrary to interruptions, the retracting phenomenon is almost always accompanied by the repetition (complete or partial) of the linguistic material and clearly causes a loss of the informational value of the retracted material, which is abandoned by the speaker in favor of the chosen alternative. As in the case of interruption, in retracting phenomena the change of prosodic envelope is again necessary. In other words, the retracting between two elements cannot be accomplished in the same prosodic envelope. Therefore, retracting is always accompanied by a prosodic break marked with the symbol "[/n]", where *n* is the number of words virtually deleted by the speaker.

²⁷ No distinction connected to possible causes of interruption is considered in this frame (e.g. the CHAT format marks when the interruption is caused by the listener). On the contrary, the format explicitly marks the distinction between *interruption* and *intentional suspension*, which frequently occurs at the end of the utterances. Intentional suspension must be marked as a generic utterance limit *"/"*.

Retracting breaks are considered a type of non terminal breaks and are highlighted only at richer levels of transcription. The symbol is inserted in the transcription after each set of fragments in the position where a restart begins. Examples of retracting are available in Appendix 1.

3.3.5.3. Retracting/interruption ambiguity

In some cases it is hard to decide whether a fragmentation phenomenon fits the definition of “retracting” or “interruption”. This can be the case when an alternative to the locution in focus is realized, but no repetition is involved. When no reason leads to the preference of one alternative, the case is treated as a retracting²⁸.

3.3.6 Summary of prosodic break types

Symbol	Description	Type of break
//	<i>Conclusive prosodic break</i>	Terminal
+	<i>Conclusive prosodic break such that the utterance is interrupted by the listener or by the speaker himself</i>	Terminal
/	<i>Non-conclusive prosodic break</i>	Non-terminal
[/n]	<i>Non-conclusive prosodic break caused by a false start</i>	Non-terminal

3.4 Procedures for prosodic tagging

The transcriptions and the prosodic tagging have been done with the following procedure:

- 1) Tagging of prosodic breaks simultaneously to the transcription by a first labeler;
- 2) Revision of tagging by a different labeler, in connection with the revision of transcripts;
- 3) Revision of tagging, in connection with the alignment, by a third labeler: after the definition of each alignment unit, the labeler always challenges the presence of a terminal break and changes the non terminal breaks only if there is a clear mistake;
- 4) Revision of the alignment and simultaneously of the prosodic tagging, with respect to the terminal breaks, by a forth labeler.

This process ensures control on the inter-annotator relevance of tags and maximum accuracy in the detection of terminal breaks. The accuracy with respect to non-terminal breaks is by definition lower. The level of inter-annotator agreement on prosodic tag assignment has been statistically evaluated twice: once before the beginning of the transcription process and before the second revision. The evaluation report is attached here in Appendix 4.

3.5 Quality assurance on prosodic tagging

In C-ORAL-BRASIL (like in C-ORAL-ROM) each position between two words is considered a possible position for a prosodic break. The prosodic tagging is based only on perceptual judgments and does not require any specific linguistic knowledge, although the notion of speech act is always familiar to the transcribers who annotated the C-ORAL-BRASIL corpus. The annotation of terminal and non-terminal breaks has been accomplished by expert transcribers. The expertise is guaranteed by the following process and the following results.

3.5.1 Transcriber training procedures

Segmenters were trained through academic classes and workshops. They were divided in two groups according to their degree of expertise. The segmenters received academic training through courses and workshops.

²⁸ In C-ORAL-ROM a supplementary sign, “[//]”, could *optionally* be used marking the fact that a probable retracting phenomenon occurs with neither partial nor complete repetition of linguistic material. The ambiguous mark is counted as a non terminal break.

Group 1: 3 members (graduate students)

Responsibilities:

1. Transcription, revision and subcorpus alignment (20 texts). This was the basis for the informational structure studies of BP.
2. Most part of the corpus alignment revision.
3. Most part of the corpus segmental revision.

Group 2: 4 members (undergraduate research assistants)

Responsibilities:

- Transcription, revision and corpus alignment (except for the subcorpus which was under the full responsibility of Group 1).

3.5.2 Prosodic segmentation validation

3.5.2.1 Previous validation

Tests were made followed by discussions. During tests, transcribers marked terminal and non-terminal prosodic boundaries in dialogue and monologue text excerpts. Texts were transcriptions not yet revised which did not carry any prosodic annotation. Transcribers worked on their own without any consultation to a third party about prosodic boundaries. The agreement between transcribers was calculated through Kappa and percentual statistics. Transcribers were expected to have a 0.8 or higher rate in terminal breaks and 0.6 or higher for non-terminal breaks. Kappa values were calculated through the *kappam.fleiss()* function in the R²⁹ computational environment. The 95% confidence interval was calculated employing the Wilson method without continuity correction³⁰.

The realistic Kappa is the agreement measurement considering not all the possible positions for break, but only the positions where at least one of the segmenters assigned a prosodic break. This test eliminates all positions where no segmenter assigned any prosodic break. This means that all the more obvious agreements are eliminated from the baseline of the agreement calculation.

The partial agreement calculates the agreement in perception of a prosodic break, no matter if terminal or non terminal; the baseline for the calculation eliminates the difference between terminal and non terminal breaks. It measures the perception of a prosodic break and not the assignment of a value to it.

3.5.2.1.1 Group 1

First test:

Segmentation of an 822 word dialogue and an 855 word monologue.

Results:

Group1 Kappa – test 1

Agreement type	total	dl	mn
General agreement	0.79	0.83	0.75
Terminal breaks	0.83	0.90	0.74

²⁹ R Development Core Team. (2010). R: A Language and Environment for Statistical Computing. Vienna. Austria: R Foundation for Statistical Computing. Retrieved from <<http://www.r-project.org>>.

³⁰ Lowry. R. (2011). VassarStats: Website for Statistical Computation. Retrieved from <<http://faculty.vassar.edu/lowry/VassarStats.html>>.

Non-terminal breaks	0.61	0.62	0.61
No breaks	0.86	0.87	0.84

Group 1 Percentage – test 1

Agreement type	dl	mn
Total agreement.....	83%	91%
Terminal breaks	16%	8%
Non-terminal breaks	6%	8%
No breaks	67%	67%
Partial agreement	10%	15%
Terminal break vs. non-terminal	3.6%	6.9%
Non-terminal break vs. no break	6.7%	8.0%
Total disagreement	1.0%	1.4%
Terminal break vs. no break	0.6%	1.1%
Terminal break vs. non-terminal vs. no break	0.4%	0.4%

Second test:

Segmentation of a 719 word dialogue and a 784 word monologue

Results:

Group 1 Kappa – test 2

Agreement type	total	dl	mn
General agreement	0.82	0.80	0.84
Terminal breaks	0.85	0.85	0.85
Non-terminal breaks	0.71	0.62	0.77
No breaks	0.88	0.87	0.89

Group 1 Percentual– test 2

Agreement type	dl	mn
Total agreement	84%	91%
Terminal breaks	15%	9%
Non-terminal breaks	7%	15%
No breaks	63%	65%
Partial agreement	13%	11%
Terminal break vs. non-terminal	5.7%	4.1%
Non-terminal break vs. no break	7.6%	6.9%
Total disagreement	0.8%	0.4%
Terminal break vs. no break	0.6%	0.1%
Terminal break vs. non-terminal vs. no break	0.3%	0.3%

Third test:

Segmentation of a 678 word dialogue and a 415 word monologue.

Results:

Group 1 Kappa – test 3

Agreement type	total	dl	mn
General agreement	0.77	0.78	0.76
Terminal breaks	0.82	0.87	0.71
Non-terminal breaks	0.62	0.58	0.66
No break	0.85	0.84	0.86

Group 1 Percentual– test 3

Agreement type	dl	mn
Total agreement	86%	93%
Terminal breaks	13%	6%
Non-terminal breaks	7%	13%
No break	65%	65%
Partial agreement	15%	16%
Terminal break vs. non-terminal	5.2%	7.5%
Non-terminal break vs. no break	9.9%	8.7%
Total disagreement	0.3%	0.5%
Terminal break vs. no break	0.1%	0.2%
Terminal break vs. non-terminal vs. no break	0.1%	0.2%

Realistic Kappa Group 1 - pre-validation

Text	overall	terminal	non-terminal
Dialogue 1	0.59	0.81	0.50
Dialogue 2	0.56	0.76	0.47
Dialogue 3	0.52	0.79	0.40
Monologue 1	0.46	0.64	0.38
Monologue 2	0.6	0.80	0.56
Monologue 3	0.45	0.63	0.37

Partial Agreement Kappa Group 1 - pre-validation

Text	Overall
Dialogue 1	0.88
Dialogue 2	0.87
Dialogue 3	0.84
Monologue 1	0.84
Monologue 2	0.89

3.5.2.1.2 Group 2

First test:

Segmentation of an 822 word dialogue and a 1014 word monologue.

Results:

Group 2 Kappa – test 1

Agreement type	total	dl	mn
General agreement	0.75	0.78	0.73
Terminal breaks	0.77	0.81	0.73
Non-terminal breaks	0.62	0.58	0.63
No break	0.83	0.86	0.80

Group 2 Percentage– test 1

Agreement type	dl	mn
General agreement	82%	78%
Terminal breaks	11%	6%
Non-terminal breaks	4%	9%
No break	67%	62%
Partial agreement	15%	20%
Terminal break vs. non-terminal	7%	7%
Non-terminal break vs. no break	8%	14%
Total disagreement	2.8%	2.2%
Terminal break vs. no break	1.6%	0.8%
Terminal break vs. non-terminal vs. no break	1.2%	1.4%

Second test:

Segmentation of a 1359 word dialogue.

Results:

Group 2 Kappa – test 2

Agreement type	dl
General agreement	0.76
Terminal breaks	0.82
Non-terminal breaks	0.57
No break	0.82

Group 2 Percentage – test 2

Agreement type	dl
Total agreement	79%
Terminal breaks	15%
Non-terminal breaks	5%
No break	59%
Partial agreement	17%
Terminal break vs. non-terminal	6%
Non-terminal break vs. no break	11%
Total disagreement	4.2%
Terminal break vs. no break	2.0%
Terminal break vs. non-terminal vs. no break	2.2%

Third test:

Segmentation of a 784 word monologue.

This test was taken by three out of four Group 2 transcribers.

Results:

Group 2 Kappa – test 3

Agreement type	mn*
General agreement	0.68
Terminal breaks	0.78
Non-terminal breaks	0.51
No break	0.75

* 3 raters

Group 2 Percentage – test 3

Agreement type	mn*
General agreement	79%
Terminal breaks	8%
Non-terminal breaks	7%
No break	64%
Partial agreement	20%
Terminal break vs. non-terminal	6%
Non-terminal break vs. no break	15%
Total disagreement	0.8%
Terminal break vs. no break	0.4%
Terminal break vs. non-terminal vs. no break	0.4%

* 3 raters

Fourth test:

Segmentation of a 681 word dialogue.

Results:

Group 2 Kappa – test 4

Agreement type	dl
General agreement	0.78
Terminal breaks	0.80
Non-terminal breaks	0.68
No break	0.84

Group 2 Percentage – test 4 (3 raters)

Agreement type	dl
Total agreement	79%
Terminal breaks	11%
Non-terminal breaks	12%
No break	55%
Partial agreement	18%
Terminal break vs. non-terminal	7%
Non-terminal break vs. no break	11%
Total disagreement	3.2%
Terminal break vs. no break	1.5%
Terminal break vs. non-terminal vs. no break	1.8%

Fifth test:

Segmentation of an 803 word dialogue.

Results:

Group 2 Kappa – test 5

Agreement type	DI
General agreement	0.77
Terminal breaks	0.85
Non-terminal breaks	0.66
No break	0.81

Group 2 Percentage – test 5

Agreement type	Mn
Total agreement	79%
Terminal breaks	12%
Non-terminal breaks	12%
No break	55%
Partial agreement	19%
Terminal break vs. non-terminal	5%

Non-terminal break vs. no break	14%
Total disagreement	2.4%
Terminal break vs. no break	0.7%
Terminal break vs. non-terminal vs. no break	1.6%

Sixth test:
Segmentation of an 1126 word monologue.

Results:

Group 2 Kappa – test 6

Agreement type	Mn
General agreement	0.77
Terminal breaks	0.82
Non-terminal breaks	0.66
No break	0.83

Group 2 Percentage – test 6

Agreement type	Mn
Total agreement.....	80%
Terminal breaks	10%
Non-terminal breaks	9%
No break	61%
Partial agreement	19%
Terminal break vs. non-terminal	6%
Non-terminal break vs. no break	13%
Total disagreement	1.3%
Terminal break vs. no break	0.4%
Terminal break vs. non-terminal vs. no break	0.9%

Seventh test:
Segmentation of a 1045 word monologue.

Results:

Group 2 Kappa– test 7

Agreement type	Mn
General agreement	0.79
Terminal breaks	0.76
Non-terminal breaks	0.70
No break	0.86

Group 2 Percentage – test 7

Agreement type	Mn
Total agreement	84%
Terminal breaks	10%
Non-terminal breaks	12%
No break	63%
Partial agreement	14%
Terminal break vs. non-terminal	5%
Non-terminal break vs. no break	8%
Total disagreement	1.9%
Terminal break vs. no break	1.3%
Terminal break vs. non-terminal vs. no break	0.6%

Eighth test
Segmentation of a 981 monologue.

Results:

Group 2 Kappa – test 8

Agreement type	Mn
General agreement	0.82
Terminal breaks	0.83
Non-terminal breaks	0.75
No break	0.87

Group 2 Percentage – test 8

Agreement type	Mn
Total agreement	87%
Terminal breaks	6%
Non-terminal breaks	8%
No break	73%
Partial agreement	12%
Terminal break vs. non-terminal	5%
Non-terminal break vs. no break	7%
Total disagreement	1.5%
Terminal break vs. no break	0.7%
Terminal break vs. non-terminal vs. no break	0.8%

Realistic Kappa Group 2 – pre-validation

Text	Overall	Terminal	Non-terminal
------	---------	----------	--------------

Dialogue 1	0.51	0.67	0.45
Dialogue 2	0.52	0.71	0.47
Dialogue 3	0.57	0.74	0.52
Dialogue 4	0.58	0.80	0.50
Monologue 1	0.47	0.64	0.44
Monologue 2*	0.38	0.72	0.23
Monologue 3	0.54	0.75	0.50
Monologue 4	0.61	0.77	0.59
Monologue 5	0.54	0.68	0.51

*3 raters

Partial Agreement Kappa Group 2 – pre-validation

Text	Overall
Dialogue 1	0.86
Dialogue 2	0.82
Dialogue 3	0.84
Dialogue 4	0.81
Monologue 1	0.80
Monologue 2*	0.75
Monologue 3	0.83
Monologue 4	0.87
Monologue 5	0.86

* 3 raters

3.5.2.2 Group 1 reevaluation and final validation

Before the beginning of the final revision of the corpus, group 1 underwent a test to reevaluate its agreement as far as prosodic boundaries annotation was concerned. This reevaluation was the final prosodic segmentation annotation for the corpus.

Test:

Segmentation of a 562 word dialogue and a 758 word monologue.

Results:

Final validation Kappa

Agreement type	total	dl	Mn
General agreement	0.86	0.86	0.85
Terminal breaks	0.87	0.87	0.86
Non-terminal breaks	0.78	0.78	0.78
No break	0.91	0.91	0.90

Final validation Percentage

Agreement type	dl	Mn
Total agreement	90.6%	90.7%
Terminal breaks	12.5%	8.3%
Non-terminal breaks	11.0%	12.4%
No break	67.1%	70.0%
Partial agreement	8.5%	9.1%
Terminal break vs. non-terminal	3.9%	3.7%
Non-terminal break vs. no break	4.6%	5.4%
Total disagreement	0.9%	0.2%
Terminal break vs. no break	0.7%	0.1%
Terminal break vs. non-terminal vs. no break	0.2%	0.1%

Realistic Kappa Group 1 – final validation

Agreement type	total	dl	mn
General agreement	0.65	0.66	0.63
Terminal breaks	0.81	0.80	0.80
Non terminal breaks	0.62	0.65	0.59

Partial agreement Kappa Group 1 – final validation

Agreement type	total	dl	mn
General agreement	0.91	0.91	0.90

Final results:

0.87 general agreement for terminal prosodic boundary annotation and 0.78 for non-terminal prosodic boundary annotation. General agreement superior to 90% for prosodic segmentation annotation.

The 95% confidence interval the C-ORAL-BRASIL prosodic segmentation convergence lies in between 0.88 and 0.99.

3.6. Alignment

Alignment is applied to the tagging of each textual string of a transcribed session with two tags corresponding to temporal information in the speech file:

- start of the alignment unit: the temporal unit corresponding to the start of the transcribed information in the speech file
- end of the alignment unit: the temporal unit of the speech file corresponding to the end of the transcribed information in the speech file

In C-ORAL-BRASIL, the alignment information is stored in an .xml file placed in the same directory of the text file and the audio file.

3.6. 1 Annotation procedure

The alignment of C-ORAL-ROM and C-ORAL-BRASIL texts is performed *after* textual transcription and prosodic tagging. The alignment of transcribed texts is achieved by an expert operator through the assistance of the WinPitch Pro alignment tool. The alignment tagging task consists in the insertion of a tag (\$) in the text after each terminal break annotated in the transcript while the audio is played (at a reduced speed).

After loading the text and the sound files to be aligned, the operator merely listens to the slow rate speech playback (between 1 and 7 times real-time), and clicks on the text segments as they are perceived, in accordance with the general choice adopted in the project in order to define a significant alignment unit (each string ending with a terminal break. See below).

Automatic dispatching of speaker turns on alignment layers is provided. The editing of segment edges is achieved through user-friendly commands using a mouse, and many other features such as the automatic scrolling of text and the dynamic adjustment of playback speed.

As out-put of this process, the system assigns two temporal units to each alignment unit :

- a) end of the alignment unit: the temporal unit of the sound file in the instant in which the tag is inserted
- b) start of the alignment unit: the temporal unit which marks the end of the previous segment

3.6.2. Prosodic tagging and the alignment unit

The Alignment of C-ORAL-BRASIL relies on two choices:

- a) Specification of the Alignment unit at utterance level (as previously defined)
- b) Rough equivalence between terminal breaks and utterance limit

Each text is aligned with respect to perceptively relevant terminal breaks annotated in the original transcripts.

When the alignment conforms to the previous requirements, the alignment file corresponds to the acoustic data base of all utterances of each speaker in the recorded session labeled with the transcribed utterance.

3.6.3 Quality assurance on the alignment

The expert operator in charge of the alignment (a fellowship undergraduate or PHD student) always considers whether the accomplished alignment unit truly corresponds, in his perception, to a speech segment ending with a terminal break. The operator may add or delete the terminal breaks annotated in the original transcripts in accordance to his personal perceptual perspective of the speech signal, thus improving the quality of the annotation.

The correspondence of an aligned segment to perceptually relevant breaks is revised immediately after the tag insertion. The same operator revises the perceptual relevance of the aligned segments of the aligned text and adjusts the edge if necessary. The alignment of overlapped strings is achieved with lower accuracy.

3.6.4 Win Pitch Pro

The alignment was performed using different versions of WinPitch Pro (www.winpitch.com). The alignment files run in a version of WinPitch Pro selected depending on the operational system of the used machine.

WinPitch Pro was chosen for different reasons: 1. Because this software was used for C-ORAL-ROM alignment; 2. Because WinPitch allows very good results with long texts alignment; 3. Because WinPitch Pro allows a friendly interface for speech analysis, specially for larger speech sequences.

3.7. PoS tagging and lemmatization

3.7.1 Tag sets

The Constraint Grammar category set of "Palavras"

<<http://beta.visl.sdu.dk/visl/pt/info/portsymbol.html>>

For more explanations and examples, as well as a comparison with the VISL form and function tags used in Palavras' graphical tree analysis, cf. [Portuguese VISL category set](http://beta.visl.sdu.dk/visl/pt/info/symbolset-manual.html) <<http://beta.visl.sdu.dk/visl/pt/info/symbolset-manual.html>>.

WORD CLASS TAGS

N Nouns

PROP Proper nouns (names)

SPEC Specifiers (defined as non-inflecting pronouns, that can't be used as prenominals):

e.g. certain indefinite pronouns, nominal quantifiers, nominal relatives

DET Determiners (defined as inflecting pronouns, that can be used as prenominals): --

e.g. articles, attributive quantifiers

PERS Personal pronouns (defined as person-inflecting pronouns)

ADJ Adjectives (including ordinals, excluding participles which are tagged V PCP)

ADV Adverbs (both 'primary' adverbs and derived adverbs ending in *-mente*)

V Verbs (full verbs, auxiliaries)

NUM Numerals (cardinals)

PRP Preposition

KS Subordinating conjunctions

KC Coordination conjunctions

IN Interjections

EC Hyphen-separated prefix ("elemento composto", category being phased out)

Where secondary tags (shown as <...>) are retained in the analysis, adverbs (ADV) and the pronoun word classes (SPEC, DET, PERS) are further differentiated into subclasses, two of which (<rel> [relatives] and <interr> [interrogatives]) often share the same word forms, and thus have to be disambiguated word-class-internally. This is in part also necessary for the <setop> [set-operator] subclass of adverbs. Other secondary tags (e.g. valency tags like <vt> for transitive verb) help disambiguate the primary [morphological and syntactical] tags, but are not disambiguated themselves. Like purely semantic tags (e.g. <prof> for profession) they may, however, be useful for resolving lexical polysemy on some higher level of analysis.

INFLECTION TAGS

Gender: M (male), F (female), M/F [for: N', PROP', SPEC', DET, PERS, ADJ, V PCP, NUM]

Number: S (singular), P (plural), S/P [for: N, PROP', SPEC', DET, PERS, ADJ, V PCP, V VFIN, INF, NUM]

Case: NOM (nominative), ACC (accusative), DAT (dative), PIV (prepositive), ACC/DAT, NOM/PIV [for: PERS]

Person: 1 (first person), 2 (second person), 3 (third person), 1S, 1P, 2S, 2P, 3S, 3P, 1/3S, 0/1/3S [for: PERS, V VFIN, V INF]

Tense: PR (present tense), IMPF (imperfecto), PS (perfeito simples), MQP (mais-que-perfeito), FUT (futuro), COND (condicional) [for: V VFIN]

Mood: IND (indicative), SUBJ (subjunctive), IMP (imperative) [for: V VFIN]

Finiteness: VFIN (finite verb), INF (infinitive), PCP (participle), GER (gerund) [for: V]

(In this table, " ' " after a word class means, that the category in question for this word class is a lexeme category, and thus derived directly from the lexicon. No " ' " means, that the category in question is a wordform category for this word class, and thus expressed by inflection.)

Derivational analysis is marked by <DERS:> (suffixation) and <DERP: ...> (prefixation), respectively.

At most intermediate levels of processing ,the asterisk marks capitalisation (<*>) and quotes (<*1> and <*2>).

\$ the dollar sign is used to mark non-word items, like punctuation marks ('\$.' for a fullstop, '\$,' for a colon) and numbers ('\$1947').

SYNTACTIC TAGS

@SUBJ> @<SUBJ subject

@ACC> @<ACC accusative (direct) object

@DAT> @<DAT dative object (only pronominal)

@PIV> @<PIV prepositional (indirect) object

@ADVS> / @SA> @<ADVS / @<SA adverbial object (place, time, duration, quantity), subject-related

@ADVO> / @OA> @<ADVO / @<OA adverbial object, object-related

@SC> @<SC subject predicative

@OC> @<OC object predicative

@ADVL> @<ADVL adverbial

@PASS> @<PASS agent of passive

(All above clause arguments [@SUBJ, @ACC, @DAT, @PIV, @ADVS, @ADVO, @SC, @OC, @PASS] and the adverbial complements [@ADVL] attach to the nearest main verb to the left [<] or right [>].)

@ADVL 'free' adverbial phrase (in non-sentence expression)

@NPHR 'free' noun phrase (in non-sentence expression without verbs)

@VOK 'vocative' (e.g. 'free' addressing proper noun in direct speech)

@>N prenominal adjective (attaches to the nearest NP-head to the right, that is not an adnominal itself)

@N< postnominal adjective (attaches to the nearest NP-head to the left, that is not an adnominal itself)

@N<PRED postnominal (in-group predicative) or predicate in small clause introduced by 'com/sem' (rare, e.g. *com a mão na bolsa, sem o pai ajudando, não conseguiu*)

@APP identifying apposition (always after NP + komma)

@>A prepositioned adverbial adjunct

(attaches to the nearest ADJ/PCP/ADV or attributive used N to the right)

@A< postpositioned adverbial adjunct (rare, e.g. *caro demais*) or dependent/argument of attributive participle (with function tag attached, e.g. @A<ADVL or @A<SC)

@PRED> 'forward' free predicative (refers to the following @SUBJ, even when this is incorporated in the VP)

@<PRED 'backward' free predicative (refers to the nearest NP-head to the left, or to the nearest @SUBJ to the left)

@P< argument of preposition
@S< sentence anaphor ('não venceu *o que* muito o contrariou')
@FAUX finite auxiliary (cp. **@#ICL-AUX**<)
@FMV finite main verb
@IAUX infinite auxiliary (cp. **@#ICL-AUX**<)
@IMV infinite main verb
@PRT-AUX< verb chain particle (preposition or "que" after auxiliary)
@CO coordinating conjunction
@SUB subordinating conjunction
@KOMP< argument of comparative (e.g. "do que" referring to *melhor*)
@COM direct comparator without preceding comparative
@PRD role predicator (e.g. "work *as*", "function *as*")
@FOC> **@<FOC** focus marker ("gosta *é* de peixe.")
@TOP topic constituent ("*Esse negócio*, não gosto dele.")
@#FS- finite subclause (combines with clausal role and intraclausal word tag, e.g. **@#FS-<ACC**
@SUB for "não acredito *que* seja verdade")
@#ICL- infinite subclause (combines with clausal role and intraclausal word tag, e.g. **@#ICL-**
SUBJ> **@IMV** in "*consertar* um relógio não é fácil")
@#ICL-AUX< argument verb in verb chain, refers to preceding auxiliary (the verb chain sequence
@FAUX - **@#ICL-AUX**< is used, where both verbs have the same subject, **@FMV** - **@#ICL-**
<ACC is used where the subjects are different)
@#AS- averbal (i.e. verb-less) subclause (combines with clausal role and intraclausal word tag, e.g.
@#AS-<ADVL @ADVL> in "*ajudou onde* possível")
@AS< argument of complementiser in averbal subclause

Below an alphabetical list of all possible syntactic tags is presented. The numbers in parentheses are statistics from 1 million words of the CETEMPúblico corpus.

@<ACC (41839) direct object
@<ACC-PASS (1476) passive use of pronoun *se*
@<ADVS and **@<ADVO** (2723) adverbial argument
@<ADVL (51664) adjunct adverbial
@<DAT (591) dative (indirect) object
@<FOC (414) focus marker (or right focus bracket)
@<OC (1740) object complement
@<PASS (1025) agent of passive
@<PIV (11601) prepositional object
@<PRED (1007) free (subject) predicative, right of main verb
@<SC (16147) subject complement
@<SUBJ (7362) subject
@>A (9037) adverbial pre-adject (intensifier before adjective, adverb, pronoun or participle)
@>N (199174) pronominal modifier
@>P (1908) modifier of prepositional phrase (intensifier, operator or focus adverb)
@>S (17) modifier of clause (intensifier, operator or focus adverb)
@A< (5257) adverbial post-adject (modifier or argument of adjective, adverb or participle)
@A<ADV (162) adverbial argument of attributive participle
@A<ADVL (1694) adverbial adjunct of attributive participle
@A<PASS (2307) agent of passive after attributive participle
@A<PIV (1220) prepositional object of attributive participle
@A<SC (38) subject complement of attributive participle
@ACC> (7904) accusative (direct) object
@ACC>-PASS (1322) passive use of pronoun *se*

@ACC>> (41) double-fronted accusative (direct) object before matrix verb
 @ADVS> and @ADVO> (365) adverbial argument
 @ADVL (6095) top node adverbial
 @ADVL> (36767) adjunct adverbial
 @ADVL>A (490) adjunct adverbial before attributive participle
 @ADVL>AS< (30) adjunct adverbial in averbal clause
 @APP (5037) identifying apposition
 @AS< (3706) clause body of averbal clause
 @CO (28420) co-ordinator
 @COM (2459) comparator (heading averbal clause)
 @DAT> (998) dative (intransitive) object
 @FAUX (13235) finite auxiliary
 @FMV (63919) finite main verb
 @FOC> (450) focus marker (or left focus bracket)
 @IAUX (3548) non-finite auxiliary
 @IMV (40677) non-finite main verb
 @KOMP< (174) argument of comparative hook
 @N< (124274) postnominal modifier or argument
 @N<PRED (10903) postnominal (in-group) predicative (or non-identifying apposition)
 @NPHR (9797) top node noun phrase
 @NUM< (188) second part of numeral chain
 @OC> (4) object complement
 @P< (154568) argument of preposition
 @PIV> (256) prepositional object
 @PRD (484) predicator (heading averbal clause)
 @PRED> (1482) free (subject) predicative, left of main verb
 @PREF (28) prefix (category being phased out)
 @PRT-AUX< (3430) auxiliary particle
 @S< (15) statement predicative (sentence apposition)
 @SC> (583) subject complement
 @SUB (11718) subordinator
 @SUBJ> (51098) subject
 @SUBJ>> (92) double-fronted subject, with interfering matrix og quoting verb
 @TOP (217) topic constituent
 @VOK (135) vocative constituent

The following combinations of the above with clausal forms occur (FS = finite subclause, ICL = non-finite clause, AS = averbal clause):

@#AS-<ADVL (1959), @#AS-<SC (28), @#AS-A< (71), @#AS-ADVL (72), @#AS-ADVL> (737), @#AS-KOMP< (648), @#AS-N< (370), @#FS-<<ACC (17), @#FS-<ACC (5835), @#FS-<ADVS (3), @#FS-<ADVL (4103), @#FS-<OC (1), @#FS-<SC (430), @#FS-<SUBJ (1102), @#FS-A< (195), @#FS-ACC> (24), @#FS-ACC>> (868), @#FS-ADVL (224), @#FS-ADVL> (1536), @#FS-AS< (17), @#FS-KOMP< (410), @#FS-N< (14393), @#FS-P< (1303), @#FS-S< (335), @#FS-SUBJ> (440), @#ICL-<ACC (3970), @#ICL-<ADVL (3232), @#ICL-<OC (23), @#ICL-<PRED (87), @#ICL-<SC (963), @#ICL-<SUBJ (860), @#ICL-ACC> (61), @#ICL-ADVL (113), @#ICL-ADVL> (478), @#ICL-APP (3), @#ICL-AS< (88), @#ICL-AUX< (16862), @#ICL-IMV (3), @#ICL-N< (3172), @#ICL-N<PRED (1101), @#ICL-NPHR (2), @#ICL-P< (11160), @#ICL-PRED> (198), @#ICL-SUBJ> (329)

SECONDARY TAGS

These are a somewhat idiosyncratic mixture of lexicon tags and mapped tags, many of them valency tags, that are used by the parser for the disambiguation of the primary tags, word class and syntactic function. Secondary tags themselves are only partly disambiguated on the syntactic level, an important example for full disambiguation being the interrogative and relative subclasses of adverbs and pronouns.

Only the most frequent secondary tags are mentioned here. Tags used in the AC/DC-project of corpus annotation are marked red.

Subclass tags

<artd> definite article (DET)
<arti> indefinite article (DET)
<quant> quantifier pronoun (DET: <quant1>, <quant2>, <quant3>, SPEC: <quant0>) or intensifier adverb
<dem> demonstrative pronoun (DET: <dem>, SPEC: <dem0>)
<poss> possessive pronoun (DET)
<refl> reflexive personal pronoun ("se" PERS ACC, "si" PERS PIV)
<si> reflexive use of 3. person possessive
<reci> reciprocal use of reflexive pronoun (= "um ao outro")
<coll> collective reflexive ("reunir-se", "associar-se")
<diff> differentiator (DET) (e.g. "e outros temas", "a mesma diferença")
<ident> identator (DET) (e.g. "o próprio usuário", "a si mesmo")
<rel> relative pronoun (DET, SPEC)
<interr> interrogative pronoun (DET, SPEC)
<post-det> typically located as post-determiner (DET @N<)
<post-attr> typically post-positioned adjective (ADJ @N<)
<ante-attr> typically pre-positioned adjective (ADJ @>N)
<adv> can be used adverbially (ADJ @ADVL)
<ks> relative adverb used like a subordinating conjunction
<kc> conjunctive adverb (pois, entretanto)
<det> determiner usage/inflection of adverb ("ela estava toda nua.")
<foc> focus marker adverb (also forms of "ser")
<prp> relative adverb used like a preposition
<KOMP> **<igual>** "equalling" comparative (ADJ, ADV) (e.g. "tanto", "tão")
<KOMP> **<corr>** correlating comparative (ADJ, ADV) (e.g. "mais velho", "melhor")
<komp> **<igual>** "equalling" particle referring to comparative (e.g. "como", "quanto")
<komp> **<corr>** "correlating" particle referring to comparative (e.g. "do=que")
<SUP> superlative
<setop> operational adverb (eg. "não", "nunca", "já", "mais" in "não mais")
<dei> discourse deictics (e.g. "aqui", "ontem")
<card> cardinal (NUM)
<NUM-ord> ordinal (ADJ)
<NUM-fract> fraction-numeral (N)
<cif> cipher (<card> NUM, <NUM-ord> ADJ)
<sam-> first part of morphologically fused word pair ("de" in "dele")
<-sam> last part of morphologically fused word pair ("ele" in "dele")
<*> 1. letter capitalized
<*1> left quote attached
<*2> right quote attached
<hyfen> hyphenated word
<ABBR> abbreviation

<prop> noun, adjective etc. used as name (upper case initial in mid-sentence)
<n> adjective or participle used as a "noun", typically as head of a nominal phrase
<fmc> finite main clause heading verb
<co-acc>, **<co-advl>**, **<co-app>**, **<co-dat>**, **<co-fmc>**, **<co-ger>**, **<co-inf>**, **<co-oc>**, **<co-pcv>**,
<co-postad>, **<co-postnom>**, **<co-pred>**, **<co-prenom>**, **<co-prparg>**, **<co-sc>**, **<co-subj>**, **<co-vfin>** co-ordinator tags indicating what is co-ordinated: @ACC, @ADVL, @APP, @DAT, main clauses, GER, INF, @OC, PCP-@IMV, @A<, @N<, @PRED, @>N, @P<, @SC, @SUBJ, VFIN (ordered list matching the <co-...> tags)

Valency tags:

<vt> monotransitive verb with accusative object
<vi> (**<ve>**) intransitive verb (ergative verb)
<vtd> ditransitive verb with accusative and dative objects
<PRP^vp> monotransitive verb with prepositional object (headed by PRP)
<PRP^vtp> ditransitive verb with accusative and prepositional objects
<vK> copula verb with subject predicative
<vtK> copula verb with object predicative
<va> transitive verb with adverbial argument:
<va+LOC>, **<va+DIR>**, **<vta+LOC>**, **<vta+DIR>**, **<vt+QUANT>** transitive verb with NP as quantitative adverbial object (e.g. "pesar")
<vt+TID> transitive verb with NP as temporal adverbial object (e.g. "durar")
<vU> "impersonal" verbs (normally in the 3S-person, e.g. "chove")
<x> auxiliary verb with infinitive (tagged @ (F)AUX - @ #ICL-AUX<)
<x+PCP> auxiliary verb with participle (tagged @ (F)AUX - @ #ICL-AUX<)
<x+GER> auxiliary verb with gerund (tagged @ (F)AUX - @ #ICL-AUX<)
<PRP^xp> auxiliary verb with (prepositional) auxiliary particle and infinitive (tagged as @ (F)AUX - @ PRT-AUX< - @ #ICL-AUX<)
<xt> "auxiliary" verb with infinitive clause subject in the accusative case, and ACI-constructions, (both tagged as @ (F)MV - @ SUBJ> - @ #ICL-ACC)
<PRP^xtp> "auxiliary" verb with accusative object and prepositional object containing an infinitive clause with its (unexpressed) subject being identical to the preceding accusative object, (tagged as @ (F)MV - @ <ACC - @ <PIV - @ #ICL-P<)
<vr> reflexive verbs (also <vrp>, <vaux-r>, <vaux-rp>)
<vq> "cognitive" verb governing a *que*-sentence
<qv> "impersonal" verb with *que*-subclause as subject predicative ("parece que", "consta que")
<+interr> "discourse" verb or nominal governing an interrogative subclause
<+n> noun governing a name (PROP) (e.g. "o senhor X")
<+num> noun governing a number (e.g. "cap. 7", "no dia 5 de dezembro")
<num+> "unit" noun (e.g. "20 metros")
<+INF> governs infinitive (N, ADJ)
<+PRP> governs prepositional phrase headed by PRP, e.g. <+sobre>
<PRP+> (typically) argument of preposition PRP
<+que> **<+PRP+que>** nominal governing a *que*-subclause (N, ADJ)

Semantic tags for nouns

PALAVRAS assigns angle-bracketed semantical tags for most nouns and verbs and some adjectives. The 157 semantic tags used for nouns are prototype classes, like <Hprof> for 'professional', which again translate into a subset of atomic features taken from a list of 16 values. The semantic tags are bilingually motivated (Portuguese-Danish translation alternatives) and polysemic words will thus have several tags. The semantical subsystem is in an experimental stage,

and not subject to a full disambiguation at the present time, though it can - together with the valency subsystem - yield a fair degree of polysemy resolution even now. The noun tag list below is in alphabetical order, with uppercase tags first.

Animal prototypes:

- <A> Animal, umbrella tag (*clone, fêmea, fóssil, parasito, predador*)
- <AA> Group of animals (*cardume, enxame, passarada, ninhada*)
- <Adom> Domestic animal or big mammal (likely to have female forms etc.: *terneiro, leão/leoa, cachorro*)
- <AAdom> Group of domestic animals (*boiada*)
- <Aich> Water-animal (*tubarão, delfim*)
- <Amyth> Mythological animal (*basilisco*)
- <Azo> Land-animal (*raposa*)
- <Aorn> Bird (*águia, bem-te-vi*)
- <Aent> Insect (*borboleta*)
- <Acell> Cell-animal (bacteria, blood cells: *linfócito*)

Plant prototypes:

- Plant, umbrella tag
- <BB> Group of plants, plantation (field, forest etc.: *mata, nabal*)
- <Btree> Tree (*oliveira, palmeira*)
- <Bflo> Flower (*rosa, taraxaco*)
- <Bbush> Bush, shrub (*rododendro, tamariz*)

cp. also <fruit> (fruit, berries, nuts: *maçã, morango, avelã, melancia*)
further proposed categories: <Bveg> (vegetable *espargo, funcho*)

Human prototypes:

- <H> Human, umbrella tag
- <HH> Group of humans (organisations, teams, companies, e.g. *editora*)
- <Hattr> Attributive human umbrella tag (many *-ista, -ante*)
- <Hbio> Human classified by biological criteria (race, age etc., *caboclo, mestiço, bebê, adulto*)
- <Hfam> Human with family or other private relation (*pai, noiva*)
- <Hideo> Ideological human (*comunista*, implies <Hattr>), also: follower, disciple (*dadaista*)
- <Hmyth> Humanoid mythical (gods, fairy tale humanoids, *curupira, duende*)
- <Hnat> Nationality human (*brasileiro, alemão*), also: inhabitant (*lisboeta*)
- <Hprof> Professional human (*marinheiro*, implies <Hattr>), also: sport, hobby (*alpinista*)
- <Hsick> Sick human (few: *asmático, diabético*, cp <sick>)
- <Htit> Title noun (*rei, senhora*)

Place and spatial prototypes:

- <L> Place, umbrella tag
- <Labs> Abstract place (*anverso. auge*)
- <Lciv> Civitas, town, country, county (equals <L> + <HH>, *cidade, país*)
- <Lcover> Cover, lid (*colcha, lona, tampa*)
- <Lh> Functional place, human built or human-used (*aeroporto, anfiteatro*, cp. <build>)

for just a building)

<**Lopening**> opening, hole (*abertura, fossa*)

<**Lpath**> Path (road, street etc.: *rua, pista*)

<**Lstar**> Star object (planets, comets: *planeta, quasar*)

<**Lsurf**> surface (*face, verniz*, cp. <**Lcover**>)

<**Ltip**> tip place, edge (*pico, pontinha*, cp. <**Labs**>)

<**Ltop**> Geographical, natural place (*promontório, pântano*)

<**Ltrap**> trap place (*armadilha, armazelo*)

<**Lwater**> Water place (river, lake, sea: *fonte, foz, lagoa*)

cp. also <**bar**> (barrier), <**build**> (building), <**inst**> (institution), <**pict**> (picture), <**sit**> (situation)

cp. also **position prototypes**: <**pos-an**> (anatomical position), <**pos-soc**> (social position)

Vehicle prototypes:

<**V**> Vehicle, umbrella tag and ground vehicle (car, train: *carro, comboio, tanque, teleférico*)

<**VV**> Group of vehicles (armada, convoy: *frota, esquadra*)

<**Vwater**> Water vehicle (ship: *navio, submersível, canoa*)

<**Vair**> Air vehicle (plane: *hidroplano, jatinho*)

Abstract prototypes:

<**ac**> Abstract countable, umbrella tag (*alternativa, chance, lazer*)

<**ac-cat**> Category word (*latinismo, número atômico*)

<**ac-sign**> sign, symbol (*parêntese, semicolcheia*)

<**am**> Abstract mass/non-countable, umbrella tag (still contains many cases that could be <**f...**>, e.g. *habilidade, legalidade*)

<**ax**> Abstract/concept, neither countable nor mass (*endogamia*), cp. <**f**>, <**sit**> etc.

cf. also <**f...**> (features), <**dir**> (direction), <**geom...**> (shapes), <**meta**> ("transparent" noun)

cf. also **concept prototypes**: <**conv**> (convention), <**domain**>, <**ism**> (ideology), <**genre**>, <**ling**> (language), <**disease**>, <**state...**>, <**therapy**>

cf. also **quantity prototypes**: <**unit**>, <**amount**>, <**cur**> (currency), <**mon**> (money amount)

Action prototypes:

<**act**> Action, umbrella tag (+CONTROL, PERFECTIVE)

<**act-beat**> beat-action (thrashing, *pancada, surra*)

<**act-d**> do-action (typically dar/fazer + N, *tentativa, teste, homenagem*)

<**act-s**> speech act or communicative act (*proposta, ordem*)

<**act-trick**> trick-action (cheat, fraud, ruse, *jeito, fraude*, similar to <**act-d**>)

<**activity**> Activity, umbrella tag (+CONTROL, IMPERFECTIVE, *correria, manejo*)

cp. also <**fight**>, <**dance**>, <**sport**>, <**game**>, <**therapy**>

Anatomical prototypes:

<an> Anatomical noun, umbrella tag (*carótida, clitoris, dorso*)
 <anmov> Movable anatomy (arm, leg, *braço, bíceps, cotovelo*)
 <anorg> Organ (heart, liver, *hipófise, coração, testículo*)
 <anost> Bone (*calcâneo, fíbula, vértebra*)
 <anzo> Animal anatomy (*rúmen, carapaça, chifres, tromba*)
 <anorn> Bird anatomy (*bico, pluma*)
 <anich> Fish anatomy (few: *brânquias, siba*)
 <anent> Insect anatomy (few: *tentáculo, olho composto*)
 <anbo> Plant anatomy (*bulbo, caule, folha*)

cp. also <f-an> (human anatomical feature)

<amount> quantity noun (*bocada, teor, sem-fim*)
 <bar> barrier noun (*dique, limite, muralha*)
 <build> building (*casa, citadela, garagem*)

Thing prototypes:

<cc> Concrete countable object, umbrella tag (*briquete, coágulo*, normally movable things, unlike <part-build>)
 <cc-h> Artifact, umbrella tag (so far empty category in PALAVRAS)
 <cc-beauty> ornamental object (few: *guirlanda, rufo*)
 <cc-board> flat long object (few: board, plank, *lousa, tabla*)
 <cc-fire> fire object (bonfire, spark, *chispa, fogo, girândola*)
 <cc-handle> handle (*garra, ansa, chupadouro*)
 <cc-light> light artifact (*lâmpião, farol, projector*)
 <cc-particle> (atomic) particle (few: *cátion, eletrônio*)
 <cc-r> read object (*carteira, cupom, bilhete, carta*, cf. <sem-r>)
 <cc-rag> cloth object (towel, napkin, carpet, rag) , cp. <mat-cloth>
 <cc-stone> (= cc-round) stones and stone-sized round objects (*pedra, itá, amonite, tijolo*)
 <cc-stick> stick object (long and thin, *vara, lança, paulito*)

cp. also <con> (container), <cord> (cord), <furn> (furniture), <pict> (picture), <tube>, <clo...> (clothing), <tool...>

Substance prototypes:

<cm> concrete mass/non-countable, umbrella tag, substance (cf. <mat>, *terra, choça, magma*)
 <cm-h> human-made substance (cf. <mat>, *cimento*)
 <cm-chem> chemical substance, also biological (*acetileno, amônio, anilina, bilirrubina*)
 <cm-gas> gas substance (so far few: *argônio*, overlap with. <cm-chem> and <cm>)
 <cm-liq> liquid substance (*azeite, gasolina, plasma*, overlap with <food> and <cm-rem>)
 <cm-rem> remedy (medical or hygiene, *antibiótico, cannabis, quinina*, part of <cm-h>, overlap with <cm-chem>)

cp. also <mat...> (materials)

Clothing prototypes:

<cloA> animal clothing (*sela, xabraqe*)
 <cloH> human clothing (*albornoz, anoraque, babadouro, bermudas*)
 <cloH-beauty> beauty clothing (e.g. jewelry, *diadema, pendente, pulseira*)
 <cloH-hat> hat (*sombrero, mitra, coroa*)
 <cloH-shoe> shoe (*bota, chinela, patim*)

Collective prototypes:

<coll> set, collective (random or systematic collection/compound/multitude of similar but distinct small parts, *conjunto, série*)
 <coll-cc> thing collective, pile (*baralho, lanço*)
 <coll-B> plant-part collective (*buquê, folhagem*)
 <coll-sem> semantic collective, collection (*arquivo, repertório*)
 <coll-tool> tool collective, set (*instrumentário, prataria*)

cp. also <HH> (group), <AA> (herd), <BB> (plantation), <VV> (convoy)

<col> colour (*amarelo, carmesim, verde-mar*)
 <con> container (implies <num+> quantifying, *ampola, xícara, aquário*)
 <conv> convention (social rule or law, *lei, preceito*)
 <cord> cord, string, rope, tape (previously <tool-tie>, *arame, fio, fibrila*)
 <cur> currency noun (countable, implies <unit>, cf. <mon>, *dirham, euro, real, dólar*)
 <dance> dance (both <activity>, <genre> and <sem-l>, *calipso, flamenco, forró*)
 <dir> direction noun (*estibordo, contrasenso, norte*)
 <domain> domain (subject matter, profession, cf. <genre>, *anatomia, citricultura, dactilografia*)
 <drink> drink (*cachaça, leite, guaraná, moca*)

Time and event prototypes:

<dur> duration noun (test: *durar*+, implies <unit>, e.g. *átimo, mês, hora*)
 cf. <per> (period) and <temp> (time point)
 <event> event (-CONTROL, PERFECTIVE, *milagre, morte*)
 cp. also <occ> (organized event), <process>, <act...> and <activity>

Feature prototypes:

<f> feature/property, umbrella tag (*problematicidade, proporcionalidade*)
 <f-an> anatomical "local" feature, includes countables, e.g. *barbela, olheiras*)
 <f-c> general countable feature (*vestígio, laivos, vinco*)
 <f-h> human physical feature, not countable (*lindura, compleição*, same as <f-phys-h>, cp. anatomical local features <f-an>)
 <f-psych> human psychological feature (*passionalidade, pavonice*, cp. passing states <state-h>)
 <f-q> quantifiable feature (e.g. *circunferência, calor*, DanGram's <f-phys> covers both <f> and <f-q>)
 <f-right> human social feature (right or duty): e.g. *copyright, privilégio, imperativo legal*)
 cp. also **state prototypes**: <state>, <state-h> (human state)

Food prototypes:

<**food**> natural/simplex food (*aveia, açúcar, carne*, so far including <spice>)
 <**food-c**> countable food (few: *ovo, dente de alho*, most are <fruit> or <food-c-h>)
 <**food-h**> human-prepared/complex culinary food (*caldo verde, lasanha*)
 <**food-c-h**> culinary countable food (*biscoito, enchido, panetone, pastel*)

cp. also <drink>, <fruit>
 further proposed categories: <spice>

<**fight**> fight, conflict (also <activity> and +TEMP, *briga, querela*)
 <**fruit**> fruit, berry, nut (still mostly marked as <food-c>, *abricote, amora, avelã, cebola*)
 <**furn**> furniture (*cama, cadeira, tambo, quadro*)

Concept prototypes:

<**game**> play, game (*bilhar, ioiô, poker*, also <activity>)
 <**genre**> genre (especially art genre, cf. <domain>, *modernismo, tropicalismo*)

cp. also <conv> (convention), <dance>, <domain>, <ism> (ideology), <ling> (language), <disease>, <sport>, <state...>, <therapy>

<**geom**> geometry noun (circle, shape, e.g. *losango, octógono, elipse*)
 <**geom-line**> line (few: *linha, percentil, curvas isobáricas*)

<**inst**> "institution", functional structure (+PLACE, +HUM, *auto-escola, bolsa, cinemateca*), cp.
 <Lh> (human-made place) and <HH> (group, organisation)
 <**ism**> ideology or other value system (*anarquismo, anti-ocidentalismo, apartheid*)
 <**ling**> language (*alemão, catalão, bengali*)
 <**mach**> machine (complex, usually with moving parts, *betoneira, embrulhador, limpa-pratos*, cp. <tool>)

<**mat**> material (*argila, bronze, granito*, cf. <cm>)
 <**mat-cloth**> cloth material (*seda, couro, vison, kevlar*), cp. <cc-rag>

<**meta**> meta noun (*tipo, espécie*)
 <**mon**> amount of money (*bolsa, custo, imposto*, cf. <cur>)
 <**month**> month noun/name (*agosto, julho*, part of <temp>)
 <**occ**> occasion, human/social event (*copa do mundo, aniversário, jantar, desfile*, cp. unorganized <event>)
 <**per**> period of time (prototypical test: *durante*, e.g. *guerra, década*, cf. <dur> and <temp>)

Part prototypes:

<**part**> distinctive or functional part (*ingrediente, parte, trecho*)
 <**part-build**> structural part of building or vehicle (*balustrada, porta, estai*)
 <**piece**> indistinctive (little) piece (*pedaço, raspa*)

cf. other structurals, such as <cc-handle>, <Ltip>

Perception prototypes:

<**percep-f**> what you feel (senses or sentiment, pain, e.g. *arrepio, aversão, desagrado, cócegas*, some overlap with <state-h>)

<percep-l> sound (what you hear, *apitadela, barrulho, berro, crepitação*)
 <percep-o> olfactory impression (what you smell, *bafo, chamuscom fragrância*)
 <percep-t> what you taste (PALAVRAS: not implemented)
 <percep-w> visual impression (what you see, *arco-iris, réstia, vislumbre*)

<pict> picture (combination of <cc>, <sem-w> and <L>, *caricatura, cintilograma, diapositivo*)

<pos-an> anatomical/body position (few: *desaprumo*)

<pos-soc> social position, job (*emprego, condado, capitania, presidência*)

<process> process (-CONTROL, -PERFECTIVE, cp. <event>, *balcanização, convecção, estagnação*)

Semantic product prototypes:

<sem> semiotic artifact, work of art, umbrella tag (all specified in PALAVRAS)

<sem-c> cognition product (concept, plan, system, *conjetura, esquema, plano, prejuízo*)

<sem-l> listen-work (music, *cantarola, prelúdio*, at the same time <genre>: *bossa nova*)

<sem-nons> nonsense, rubbish (implies <sem-s>, *galimatias, farelório*)

<sem-r> read-work (*biografia, dissertação, e-mail, ficha cadastral*)

<sem-s> speak-work (*palestra, piada, exposto*)

<sem-w> watch-work (*filme, esquete, mininovela*)

cp. <ac-s> (speech act), <talk>

cf. also **concept prototypes**: <conv> (convention), <domain>, <ism> (ideology), <game>, <genre>, <ling> (language), <disease>, <state...>, <therapy>

<sick> disease (*acne, AIDS, sida, alcoolismo*, cp. <Hsick>)

<sick-c> countable disease-object (*abscesso, berruga, cicatriz, gangrena*)

State-of-affairs prototypes:

<sit> psychological situation or physical state of affairs (*reclusão, arruaça, ilegalidade*, more complex and more "locative" than <state> and <state-h>)

<state> state (of something, otherwise <sit>), *abundância, calma, baixa-mar, equilíbrio*

<state-h> human state (*desamparo, desesperança, dormência, euforia, febre*, cp. <f-psych> and <f-phys-h>, which cover innate features)

<sport> sport (*capoeira, futebol, golfe*, also <activity> and <domain>)

<talk> speech situation, talk, discussion, quarrel (implies <activity> and <sd>, *entrevista, lero-lero*)

<temp> temporal object, point in time (*amanhecer, novilúnio*, test: *até+*, cf. <dur> and <per>)

<therapy> therapy (also <domain> and <activity>, *acupuntura, balneoterapia*)

Tool prototypes:

<tool> tool, umbrella tag (*abana-moscas, lápis, computador, maceta*, "handable", cf. <mach>)

<tool-cut> cutting tool, knife (*canivete, espada*)

<**tool-gun**> shooting tool, gun (*carabina, metralhadora, helicanão*, in Dangram: <tool-shoot>)
<**tool-mus**> musical instrument (*clavicórdio, ocarina, violão*)
<**tool-sail**> sailing tool, sail (*vela latina, joanete, coringa*)

cp. also <mach> (machine)

<**tube**> tube object (*cânula, gasoduto, zarabatana*, shape-category, typically with another category, like <an> or <tool>)

<**unit**> unit noun (always implying <num+>, implied by <cur> and <dur>, e.g. *caloria, centímetro, lúmen*))

Weather prototypes:

<**wea**> weather (states), umbrella tag (*friagem, bruma*)
<**wea-c**> countable weather phenomenon (*nuvem, tsunami*)
<**wea-rain**> rain and other precipitation (*chuvisco, tromba d'água, granizo*)
<**wea-wind**> wind, storm (*brisa, furacão*)

SEMANTIC TAGS FOR PROPER NOUNS (HAREM tags in parenthesis)

Person categories <hum>, HAREM PESSOA:

<**hum**> (INDIVIDUAL) person name (cp. <H>)
<**official**> (CARGO) official function (~ cp. <Htitle> and <Hprof>)
<**member**> (MEMBRO) member

Organisation/Group categories <org>, HAREM ORGANIZACAO:

<**admin**> (ADMINISTRACAO, ORG.) administrative body (government, town administration etc.)
<**org**> (INSTITUICAO/EMPRESA) commercial or non-commercial, non-administrative non-party organisations (not place-bound, therefore not the same as <Linst>)
<**inst**> (EMPRESA) organized site (e.g. restaurant, cp. <Linst>)
<**media**> (EMPRESA) media organisation (e.g. newspaper, tv channel)
<**party**> (INSTITUICAO) political party
<**suborg**> (SUB) organized part of any of the above

currently unsupported: <**company**> (EMPRESA) company (not site-bound, unlike <inst>, now fused with. <org>)

Group categories, HAREM PESSOA:

<**groupind**> (GROUPOIND) people, family
<**groupofficial**> (GROUPOCARGO) board, government (not fully implemented)

currently unsupported: <**grouporg**> (GROUPOMEMBRO) club, e.g. football club (now fused with <org>)

Place categories <top>, HAREM LOCAL:

<top> (GEOGRAFICO) geographical location (cp. <Ltop>)
<civ> (ADMINISTRACAO, LOC.) civitas (country, town, state, cp. <Lciv>)
<address> (CORREIO) address (including numbers etc.)
<site> (ALARGADO) functional place (cp. <Lh>)
<virtual> (VIRTUAL) virtual place
<astro> (OBJECTO) astronomical place (in HAREM object, not place)

suggested: <road> (ALARGADO) roads, motorway (unlike <address>)

Event categories <occ>, HAREM ACONTECIMENTO:

<occ> (ORGANIZADO) organised event
<event> (EVENTO) non-organised event
<history> (EFEMERIDE) one-time [historical] occurrence

Work of art/product categories <tit>, HAREM OBRA:

<tit> (REPRODUZIDO) [title of] reproduced work, copy
<pub> (PUBLICACAO) [scientific] publication
<product> (PRODUTO) product brand
<V> (PRODUTO) vehicle brand (cp. <V>, <Vair>, <Vwater>)
<artwork> (ARTE) work of art

Abstract categories <brand>, HAREM ABSTRACCAO:

<brand> (MARCA) brand
<genre> (DISCIPLINA) subject matter
<school> (ESCOLA) school of thought
<idea> (IDEA) idea, concept
<plan> (PLANO) named plan, project
<author> (OBRA) artist's name, standing for body of work
<absname> (NOME)
<disease> (ESTADO) physiological state, in particular: disease

Thing categories <common>, HAREM COISA:

<object> (OBJECT) named object
<common> (OBJECT) common noun used as name
<mat> (SUBSTANCIA) substance
<class> (CLASSE) classification category for things
<plant> (CLASSE) plant name
<currency> (MOEDA) currency name (also marked on the number)

Time categories (if used for NUM or N rather than PROP, marked only on the numeral or noun, without MWE'ing, unlike **HAREM TEMPO**):

<date> (DATA) date
<hour> (HORA) hour
<period> (PERIODO) period
<cyclic> (CICLICO) cyclic time expression

Numeric value categories (marked only on the numeral, without MWE'ing, unlike **HAREM VALOR**):

- <quantity> (QUANTIDADE) simple measuring numeral
- <prednum> (CLASSIFICADO) predicating numeral
- <currency> (MOEDA) currency name (also marked on the unit)

OTHER SEMANTICALLY INSPIRED CATEGORIES

<mass> mass noun (e.g. "leite", "a'gua")

<jh> adjective modifying human noun

<jn> adjective modifying inanimate noun

<ja> adjective modifying animal

<jb> adjective modifying plant

<col> color adjective

<nat> nationality adjective (also: from a certain town etc.)

<attr> (human) attributive adjective (not fully implemented, cp. <Hattr>, e.g. "um presidente COMUNISTA")

<vH> verb with human subject

<vN> verb with inanimate subject

3.7.2 automatic PoS tagging: tool and evaluation

The Morphosyntactic Tagging of the C-ORAL-Brasil Corpus

Eckhard Bick

Institute of Language and Communication

University of Southern Denmark

eckhard.bick@mail.dk

3.7.2.1 Introduction

The following is a technical description of the morphosyntactic tagging of the corpus, but also of some of the problems encountered and solutions applied during the annotation process. Usually, a corpus publisher will use a parser *as is*, maybe provide some manual training data for optimization, but in the case of the C-ORAL-Brasil corpus grammatical annotation was a separate sub-project, and the parser used (PALAVRAS) was not run as a black box, but actively adapted to speech data at both the lexical and syntactic levels. The annotation is a higher-level annotation in the sense that it does not only provide part of speech, orthographical normalization and a morphological analysis, but also tags for syntactic function and (optionally) dependency links, as well as some so-called secondary tags for semantic class.

Using automatic annotation, either on its own or as a pre-step for manual revision, is an obvious choice for a corpus this size (~ 300.000 words). Thus, previous European C-ORAL sister projects employed statistical part of speech taggers for this task, such as the PiTagger system (Moneglia et al 204) for the Italian section, which had access to a lexicon-based analyzer, a standard lexicon (107.00 lemmas), a training corpus (50.000 words) and a special pre-dictionary covering about 2000 non-standard and dialectal forms. For the European Portuguese section, the Brill tagger (Brill 1993) was used, trained on a written Portuguese corpus of 250.000 words. While no higher-level, syntactic

annotation was attempted in the European C-ORAL, other speech corpus projects have opted for full treebank annotation, such as the Arabic treebank describe by Maamouri et al. (2010), which combined manual selection of analyzer suggestion, followed by an automatic syntactic parsing stage. However, the Arabic treebank was built from broadcast data, not interviews or spontaneous dialogue, so no direct comparison can be made with C-ORAL, given the much lower need for word form standardization and discourse meta annotation in the transcription of news feed data.

For our own work we used the Palavras parser (Bick 2000) as a point of departure. Palavras is a Constraint Grammar (CG) parser that is mostly used for the annotation of written data, but has demonstrated great robustness in the face of genre variation (as, for instance, in the Linguateca³¹ project and the CorpusEye corpora³²). With lexical adaptation and various filter programs, the parser has also been used for non-standard language varieties, such as historical texts (Bick & Módolo 2005). The Constraint Grammar paradigm (Karlsson 1995), which the Palavras parser adheres to, can be described as a dualism of a robust, modular disambiguation methodology for Natural Language Processing (NLP) on the one hand, and a linguistic-descriptive convention on the other hand, encoding linguistic analyses as token-based tags and function-mediated dependency structures. Both the method and the descriptive tradition offer a number of formal advantages for the annotation of non-standard language data such as speech. First, because CG systems have a modular architecture with a clear separation of lexica, analyzers and grammars (rule sets) for successive levels of analysis, it is relatively easy to add specialized lexica or morphological filters, as well as add specific grammar modules. Second, CG's token-based annotation, where even higher-level structural information is strictly token-based, allows a corpus project to maintain several layers of annotation in parallel (such as discourse markers as opposed to clause boundaries), even allowing rules handling one layer to make reference to tags from another layer. Several speech annotation projects have made use of these advantages, such as Müürisep & Uiho (2006) for Estonian and Bondi et al. (2009) for the Nordic Dialect Corpus, though the latter used a hybrid technique, where written-text CG was used to annotate a chunk of speech data from the Oslo area, which was then manually corrected and used to train a Decision Tree Tagger (Schmid 1994) for use on other (Norwegian) dialects. In the European C-ORAL context, the Spanish section employed CG-inspired rules for part-of-speech disambiguation of morphological output from the GRAMPAL system (Moreno 2003), and for the Palavras parser itself (Bick (1998), early experiments with a Constraint-Grammar-only solution were carried out in connection with the morphosyntactic annotation of the Brazilian NURC corpus (“Norma Lingüística Urbana Culta”, Castilho 1993).

3.8 Tagging format and categories

3.8.1 Two-level annotation

C-ORAL-Brasil uses a number of symbols and encoding conventions to handle data flow issues like turn taking, prosodic breaks, speaker overlap, retractions and interruptions. Such encoding is either in non-alphanumeric form (<, /, +), or not part of an utterance (speaker names), so they either cannot or must not be analyzed by the parser. To both maintain this meta-information and to provide text-only input to the parser, we opted for a two-level annotation, where meta-information is “stored” in angle brackets on separate lines as corpus meta markup, reminiscent of e.g. <source>, <s> and <p> markers in written corpus annotation. PALAVRAS' annotation is transparent to such markup and will not change, remove or try to analyze it. Consider the following two-turn example, first in C-ORAL native annotation, then in vertical CG format, after parsing.

*LEO: o Juninho <foi> //

*GIL: <ô / mas> / voltando à questão / **falando em** [/2] e também falando em povo mascarado / esse povo do Galáticos é muito palha / eu acho que es nũ deviam mais participar / e <tal> //

31 www.linguateca.pt

32 www.corp.hum.sdu.dk

<LEO:>
 o [o] <artd> DET M S @>N
 Juninho [Juninho] <hum> <newlex> <*> PROP M S
 @SUBJ>
 <overlap-start>
 foi [ser] <fmc> V PS 3S IND VFIN @FMV
 <overlap-stop>
 \$;
 <GIL:>
 <overlap-start>
 ô [ô] <newlex> IN @ADVL
 \$,
 mas [mas] KC
 <overlap-stop>
 \$,
 voltando [voltar] V GER @IMV @#ICL-ADVL>
 a [a] <sam-> PRP @<PIV
 a [o] <-sam> <artd> DET F S @>N
 questão [questão] <ac> N F S @P<
 \$,
 <retract:falando_em>
 e [e] KC
 também [também] ADV @ADVL>
 falando [falar] <vH> V GER @IMV @#ICL-<ADVL
 em [em] PRP @<PIV
 povo [povo] <HH> N M S @P<
 mascarado [mascarar] <vH> V PCP M S @N<

\$,
 esse [esse] <dem> DET M S @>N
 povo [povo] <HH> N M S @SUBJ>
 de [de] <sam-> PRP @N<
 o [o] <-sam> <artd> DET M S @>N
 Galáticos [Galáticos] <org> <newlex> <*>
 PROP M P @P<
 é [ser] <vK> <fmc> V PR 3S IND VFIN @FMV
 muito [muito] <quant> ADV @<ADVL
 palha [palha] <cm> N F S @<SC
 \$,
 eu [eu] PERS M/F 1S NOM @SUBJ>
 acho [achar] <vH> <fmc> V PR 1S IND VFIN
 @FMV
 que [que] KS @SUB @#FS-<ACC
 es OALT eles [eles] PERS M 3P NOM @SUBJ>
 nã OALT não [não] ADV @<ADVL
 deviam [dever] V IMPF 3P IND VFIN @FAUX
 mais [mais] ADV @<ADVL
 participar [participar] <vH> V INF @IMV
 @#ICL-AUX<
 \$,
 e [e] KC
 <overlap-start>
 tal [tal] <diff> <KOMP> DET M/F S @<OC
 <overlap-stop>
 \$;

Here, only lines not starting in '<' are part of the morphosyntactic annotation. Speaker names are separate meta tags <GIL:>, and overlaps (<....>) are marked with <overlap-start> and <overlap-stop> markers. It is a clear advantage for the parser that retractions are pre-marked manually in brackets at the start point of the retraction, providing the precise number of retracted words. Our preprocessor module only needs to eliminate the words in question from the surface level to enable much smoother syntactic parses. Word repetitions or self-corrections, if allowed to persist at the surface level, would be problematic for CG rules at all levels, interfering not only with the implementation of linguistic universals like the uniqueness principle, but also with word class adjacency and agreement rules.

As can be seen from the example, the surface-deleted words will be stored in a special <retract:...> tag, maintaining the principle of two-level annotation, where two levels of annotation are separated, but not mutually exclusive. The same procedure is used for so-called non-words, which come in 2 types - first, a few non-word surface strings without special markup ('hhh' and 'xxx'), and second, incomplete words (contractions), which are marked with an initial &-sign.

*GIL: **hhh** eu tenho &**dire**

<GIL:>
 <nonword:hhh>
 eu [eu] PERS M/F 1S NOM @SUBJ>
 tenho [ter] <fmc> V PR 1S IND VFIN @FMV
 <nonword:&dire>

A special complication arose from the fact that overlap and retraction markings can be nested and/or overlapping, as the example below shows, requiring careful ordering of string matches, for instance to prevent retractions from getting “invisible” within (de-texted) speaker overlap markers. Also, since overlaps and non-words can appear within the scope of a retraction, they would change the latter's word count if removed too early, and possibly affect real words further to the left.

*GIL: <eu &a [/2] eu acho que é> esse [/2] é esse aqui o' // <&he> +

<GIL:> <overlap-start> <retract:eu_&a> eu [eu] PERS M/F 1S NOM @SUBJ> acho [achar] <vH> V PR 1S IND VFIN @FMV que [que] KS @SUB @#FS-<ACC <retract:é> <overlap-stop> <retract:esse>	é [ser] <vK> V PR 3S IND VFIN @FMV esse [esse] <dem> DET M S @<SC aqui [aqui] ADV @N< o' OALT olha [olhar] <vH> V PR 3S IND VFIN @FMV \$; <overlap-start> <nonword:&he> <overlap-stop> \$...
---	--

It should be noted that non-inclusive bracketing overlaps of the type <a> represent a general annotation problem, even for elaborate xml encoding schemes, since the latter do not envision non-projective (overlapping) tree structures, so the CG annotation chosen here can be said to be a fairly robust solution.

3.8.2 Morphosyntactic tag fields

As can be seen from the example, the morphosyntactic annotation itself consists of token-based tags ordered into fields:

Word form	lemma (base form)	optional: secondary tags	part of speech	inflexion tags	syntactic function	dependency link
acho	[achar]	<vH>	V	PR 1S IND VFIN	@FMV	#n->m
		human verb	verb	present tense 1 person singular indicative finite verb	finite main verb	n=self id m=mother id

Table 1: Tag fields

3.8.2.1 Word form

The word form field contains the original corpus token *as is*, optionally followed by a normalized form. The parser's own normalizations (ou/oi variation, European-Brazilian Portuguese, some heuristics) are marked ALT, normalizations drawn from the special C-ORAL-Brasil lexicon are marked OALT:

o' OALT olha [olhar] <vH> V PR 3S IND VFIN @FMV

3.8.2.2 Lemma / Base form

PALAVRAS uses a linguistic morphological analyzer, not a fullform lexicon, cutting off inflexion endings and affixes until a lexicon-registered base form is matched, that fulfills all possible combinatorial conditions. Normally base forms are lemmas, but the term base form implies that derivational or compound analysis may be performed where simple inflectional analysis does not suffice.

3.8.2.3 Secondary tags

Whereas primary tags, such as part of speech and inflexion, are obligatory fields and will be disambiguated by PALAVRAS, secondary tags cover additional lexical or functional information that is optional and not necessarily disambiguated. The parser uses secondary tags such as valency and semantic class tags to provide better context for the assignment disambiguation of primary tags. Only certain secondary tags have been retained in the C-ORAL-Brasil annotation - in particular,

subclass categories such as <rel> relative, <interr> interrogative, <art> article, which may be needed to filter the parser's own PoS categories into other category sets according to corpus user preferences. In order to allow semanticized searches, frequency generalizations and the extraction of selection restrictions or metaphor candidates, semantic class tags are also shown, comprising some 200 tags for nouns, and the +Hum tag for verbs (<vH>).

The secondary tag field also contains a few tags for orthographical features:

<*>	upper case
<*1>	left quote
<*2>	right quote
<sam->	first part of contraction
<-sam>	second part of contraction
<pp>	complex adverb (prepositional phrase)
<hyfen>	hyphenated word
<newlex>	from add-on lexicon

3.8.2.4 Part of speech

PALAVRAS uses, wherever possible, a purely morphological definition of word classes, trying to keep linguistic form and function separate. Thus, nouns (N) are defined as a word class where gender is a lexeme category (i.e. un-inflecting) and number a word form category (i.e. inflecting), while both are word form categories for adjectives (ADJ), and both are lexeme categories for proper nouns (PROP). Similarly, nominal pronouns are subdivided into gender-number-inflecting determiners (DET), uninflecting independent pronouns (so-called specifiers, SPEC) and personal pronouns (PERS), characterized by the categories person, number and case. Pronoun categories that traditionally are defined syntactically-functionally (relatives, reflexives) or semantically (quantifier indefinites, possessives, demonstratives) are subcategorized with secondary tags (<rel>, <interr>, <refl>, <dem>, <poss>). From a morphological and word form point of view, Portuguese articles are indistinguishable from determiner pronouns, and are therefore also marked with secondary tags (<artd> DET and <arti> DET).

Non-inflecting word classes are abbreviated as follows:

ADV	adverb
NUM	numeral
PRP	preposition
KC	coordinating conjunction
KS	subordinating conjunction
IN	interjection

Finally, PREF is used for non-word prefixes occurring in isolation.

3.8.2.5 Morphology / inflexion

The inflectional potential of Portuguese word classes can be seen in the following overview, where '+' means word form category (inflecting) and '+*' means lexeme category (non-inflecting).

	gender	number	case	person	tense	mode
	M=male F=female	S=singular P=plural	NOM=nominative ACC=accusative DAT=dative PIV=prepositive	1=1. person 2=2. person 3=3. person	PR=present IMPF =imperfect PS=perfeito simples FUT=future	IND=indicative SUBJ =subjunctive IMP=imperative COND =conditional

N	+	*	+			
PROP	+	*	+	*		
ADJ	+		+			
PERS	+		+		+	
DET	+		+			
SPEC	M*		S*			
V VFIN					+	+
V INF					+	
V PCP	+		+			
V GER						

Table 2: Inflection categories and part of speech

As can be seen, DET and PCP (participle) are, from an inflectional point of view, "adjectival" classes, while gerunds (GER) are the only non-inflecting verbal class, matching the class' adverbial usage characteristics.

3.8.2.6 Syntactic function

The CG syntactic tags employed here contain both functional and shallow dependency information. The tag field is marked by a '@' prefix, and multiple tags are possible for the same token in the case of unresolved ambiguity. Dependency is expressed by so-called shallow dependency markers expressed as '>' (right) and '<' (left) attachment markers, indicating in which direction the head of the token in question can be found. Very rarely, double arrows (>>, <<) are used to indicate a long distance attachment, beyond the first qualifying head. Head form information is provided at the arrow point, syntactic function at the arrow base. Usually, head form is left implicit at the clause level (e.g. @<SUBJ for right-positioned subject), because the head will always be the verb. At the group level, the inverse is true, function is often omitted and implied by head type. Thus, @>N means pre-nominal dependent, and @N< post-nominal dependent, and the presence of either will define an np constituent, even if the head is not a noun. For instance, the functional force of a dependent article/dependent may project "noun-hood" on an adjectival head, as in *os velhos, um doente*.

In classical PALAVRAS annotation, subclause function is tagged as an "external", second tag on subordinators (finite subclauses) or verbs (non-finite subclauses), prefixing either @FS- (finite subclause), @ICL- (non-finite subclause) or @AS- (averbal/elliptic subclause). Thus, @FS-N< means a (postnominal) relative clause. At the same time, the word in question will retain its "internal" tag, e.g. @SUBJ> or @ACC> for a relative pronoun, or @IAUX (non-finite auxiliary) for a verb. When filtered to international VISL annotation, the external function tag will always be on the first verb, turning the auxiliary or main verb tag into a secondary tag (<aux>, <mv>), .

The following syntactic tags are used:

Valency-bound verb arguments

- @SUBJ> @<SUBJ subject
- @ACC> @<ACC accusative (direct) object
- @DAT> @<DAT dative object (only pronominal)
- @PIV> @<PIV prepositional (indirect) object
- @ADVS>/@SA> @<ADVS/@<SA adverbial predicative , equivalent to nominal @SC

@ADVO>/@OA> @<ADVO/@<OA adverbial object predicative, matches nominal @OC
 @SC> @<SC subject predicative
 @OC> @<OC object predicative

Free clause constituents

@ADVL> @<ADVL adverbial
 @PRED> 'forward' free predicative, @<PRED 'backward' free predicative
 @<PASS agent of passive
 @ADVL 'free' adverbial phrase (in non-sentence expression)
 @NPHR 'free' noun phrase (in non-sentence expression without verbs)
 @VOK 'vocative' (e.g. 'free' addressing proper noun in direct speech)
 @S< sentence-related "apposition"

Noun phrase:

@>N prenominal adjunct
 @N< postnominal adjunct
 @N<PRED postnominal (in-group predicative)
 @APP identifying apposition

Adjective, determiner and adverb phrases:

@>A prepositioned adverbial adjunct
 @A< postpositioned adverbial adjunct

Prepositional, comparative and headed averbal phrases:

@P< argument of preposition
 @KOMP< argument of comparative (e.g. "do que" referring to *melhor*)
 @AS< argument of complementiser in averbal subclause

Verbal constituent:

@FAUX finite auxiliary (cp. @#ICL-AUX<)
 @FMV finite main verb
 @IAUX infinite auxiliary (cp. @#ICL-AUX<)
 @IMV infinite main verb
 @PRT-AUX< verb chain particle (preposition or "que" after auxiliary)
 @#ICL-AUX< argument verb in verb chain, refers to preceding auxiliary

Coordination and subordination:

@CO coordinating conjunction
 @SUB subordinating conjunction
 @COM direct comparator without comparative
 @PRD role predicator (e.g. "work *as*")

Focus and topic markers

@FOC> @<FOC focus marker ("gosta *é* de peixe.")
 @TOP topic constituent ("*Esse negócio*, não gosto dele.")

3.8.2.7 Dependency links and syntactic trees

On top of the shallow dependency direction markers that are part of the syntactic @-tags, PALAVRAS can provide explicit, numbered dependency arcs, i.e. complete dependency trees. The dependency field has the form #n->m, where 'n' is the ID of the token in question, 'm' the ID of its head, as shown below for the sentence *O último diagnóstico elaborado pela Comissão Nacional não deixa dúvidas*:

O <artd>	DET M S	@>N	#1->3
último	ADJ M S	@>N	#2->3

diagnóstico	N M S	@SUBJ>	#3->9
elaborado	V PCP2 M S	@IMV @#ICL-N<	#4->3
por	PRP	@<PASS	#5->4
a <artd>	DET F S	@>N	#6->7
Comissão=Nacional	PROP F S	@P<	#7->5
não	ADV	@ADVL>	#8->9
deixa	V PR 3S	@FMV	#9->0
dúvidas	N F P	@<ACC	#10->9
\$.			#11->0

Syntactic tree structures like these can be transformed into a variety of formats. For complete trees, PALAVRAS offers, for instance, vertical VISL indentation trees, traditional Chomskyan constituent brackets, PENN treebank annotation, TIGER treebank xml and MALT dependency xml markup. Visualizers exist for each of these formats, such as the interactive graphical tree-builder for the VISL format:

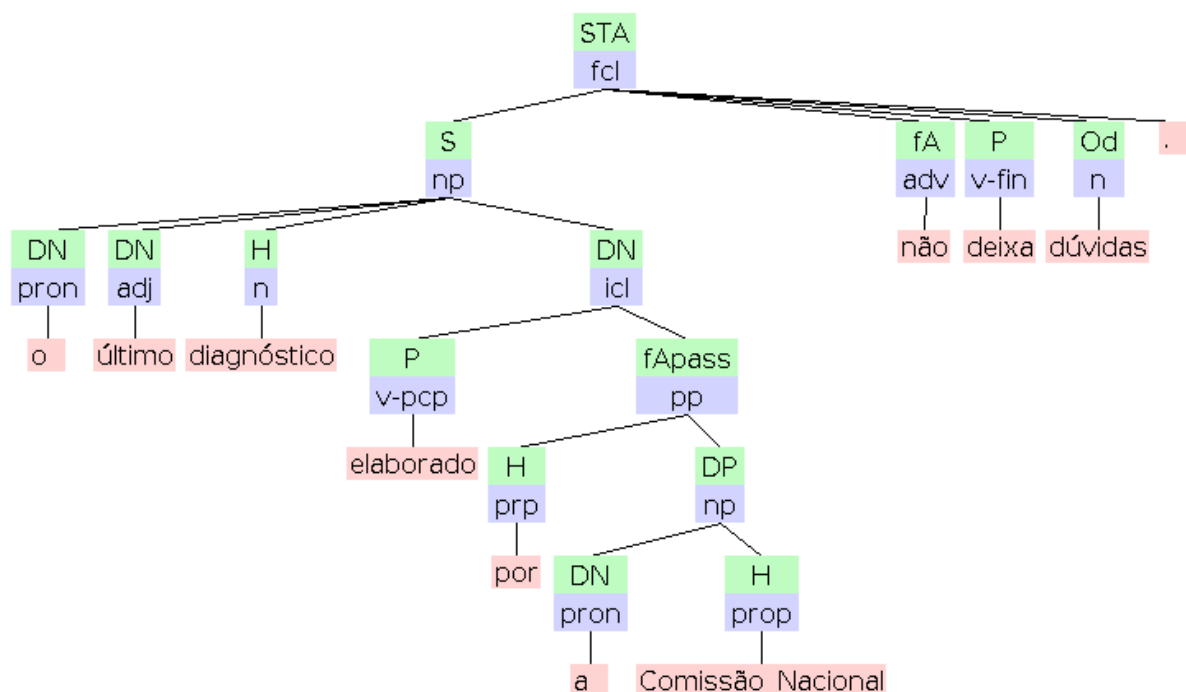


Fig. 1: Graphical syntactic tree structures

3.9 The tool

Technically, the Palavras parser consists of a chain of Constraint Grammar rules, bundled in sets of increasing heuristicity, successively handling ever higher (deeper) levels of analysis, progressing from morphological disambiguation and PoS tagging, over syntactic function mapping and dependency relations, to semantic role annotation, Named Entity Recognition and application-oriented modules. Input to this chain of grammars is provided by a preprocessor/tokenizer and a morphological analyzer program supported by large lexica covering inflexional paradigms, valency potential, semantic class ontologies etc. All lexical information is encoded, CG-style, as token-linked tags on reading lines. Ambiguous reading lines for a given word are called a *cohort*, as would be the case for the typical verbo-nominal Portuguese ambiguities involving -a and -o endings:

"<casa> "
 "casa" <build> N F S ('house')
 "casar" <vt> <vi> <com^vp> <vr> <com^vrp> <vH> V IMP 2S VFIN ('marry!')
 "casar" <vt> <vi> <com^vp> <vr> <com^vrp> <vH> V PR 3S IND VFIN ('marries')
 "<acordo ALT acordo> "
 "acordo" <sem-c> <+com> <+sobre> <+entre> <de+> <+n> ('deal, contract')
 "acordar" <ve> <vt> <vK> V PR 1S IND VFIN ('wake up')

A distinction is made between primary tags, which are slated for disambiguation (e.g. N, V), and secondary tags that are not (or not at this level) intended for disambiguation (<...> tags), but rather to provide contextual clues for CG rules in the process of primary disambiguation. Thus, a transitive verb tag <vt> and a human noun tag <H> may help to assign subject and object functions to the nouns in a sentence.

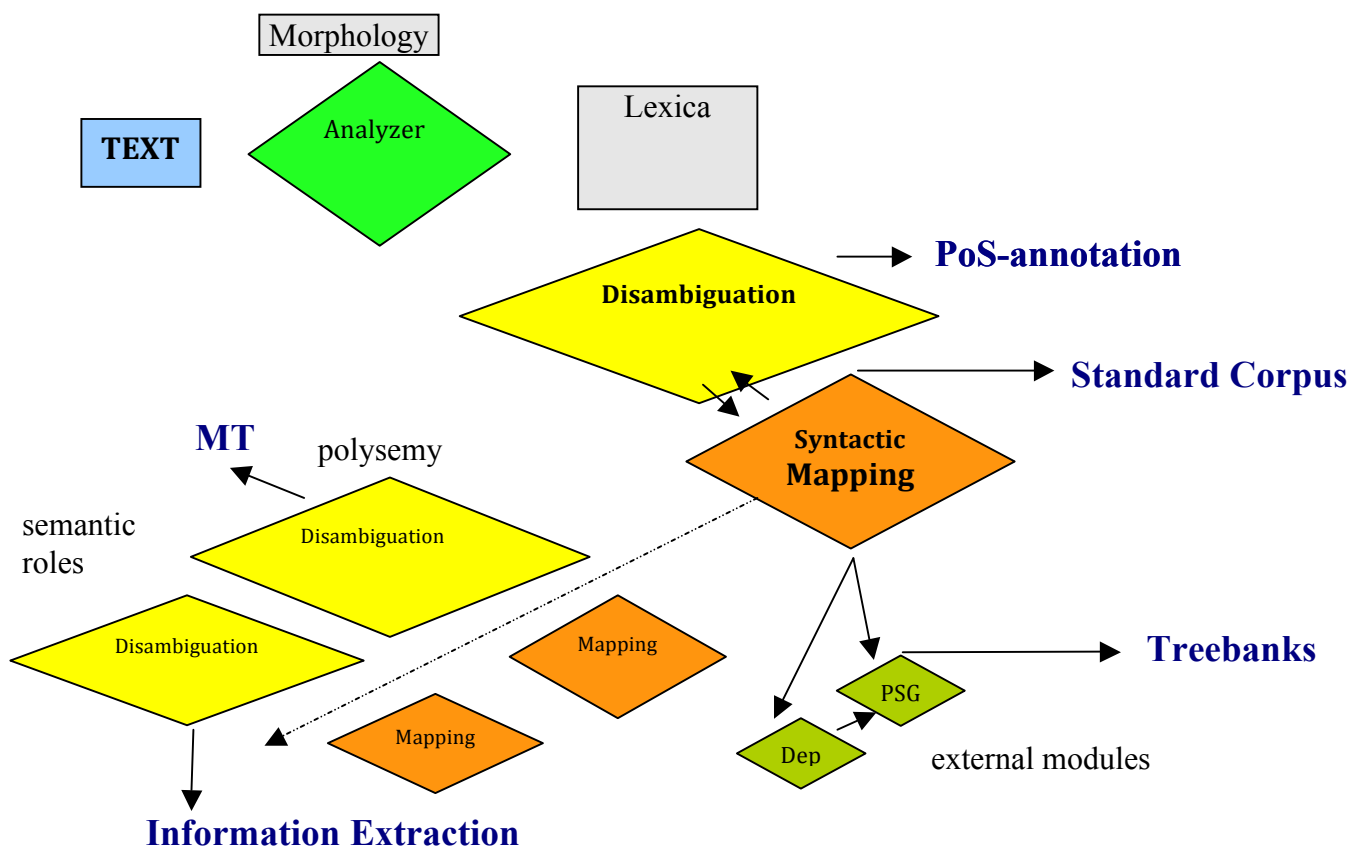


Fig. 2: Parser flow chart

PALAVRAS uses about 6.000 contextual CG rules that either remove, select, add, map or substitute tags/readings. Apart from other tags (associated with any other word in the sentence), the CG formalism allows rules to make reference to word- or reading-related statistical-numerical information, match regular expressions in word forms, tags and lexemes, or unify category features across constituents. Though somewhat foreign to the formalism's reductionist methodology, even generative rewriting rules can be expressed in Constraint Grammar, with the help of so-called templates.

Most rules, however, are fairly straight forward disambiguation rules such as the following, that removes a finite verb reading (VFIN), if there is a safe (C) preposition reading to the left:

REMOVE VFIN IF (-1C PRP) ; # remove a

While such close context could be captured by probabilistic Hidden Markov Model (HMM) n-gram modeling, too, many CG rules have global sentence scope and are considerably more powerful. Consider for instance the following uniqueness rule, saying that a word cannot be a finite verb if there already is a finite verb anywhere (*) to the left (*-1), without a clause boundary (CLB) or coordinator (KC) in between (BARRIER).

REMOVE VFIN (*-1 VFIN BARRIER CLB OR KC)

Given the architecture and rule methodology of the parser, three challenges can be identified with regard to its application to oral data, affecting lexical recall on the one hand (2) and contextual disambiguation on the other (1,3). In many ways, the problems are similar to the ones encountered in the annotation of historical language data (Bick & Módolo 2005).

- How to maintain corpus meta information from the non-grammatical annotation layers, while still providing “running text” input to the parser and its analyzer
- How to adapt the lexicon and/or to change the word forms to allow input to be recognized as ordinary written, modern, standard Brazilian Portuguese, while at the same time maintaining the oral transcription forms
- How to provide syntactic breaks, in the absence of ordinary punctuation, to allow the definition of delimited windows for contextual disambiguation

3.10 Tokenization

The simplest automatic tokenization would define a non-punctuation token as any string of alphanumericals surrounded by spaces or punctuation, needing linguistic rules only for the recognition of abbreviations and numerical expressions. However, since PALAVRAS is not just a part of speech tagger, but a syntactic parser providing deep tree analyses, it performs both fusions and splittings of what otherwise would have been tokens, in order to facilitate the recognition and assignment of syntactic form and function.

3.10.1 Multi-word expressions

Multi-word expressions are fused with equal-signs (=):

Names: São=Paulo, Antônio Carlos

complex prepositions: em=vez=de, em=cima=de

complex conjunctions: do=que

Pronouns: todo=mundo, cada=um

complex adverbs: hoje=em=dia, pelo=amor=de=Deus

Interjections: hum=hum, ham=ham

PALAVRAS has named-entity recognition (NER) as one of its specialities³³, and systematically fuses the individual parts of names in order to assign a semantic classification and syntactic function to the whole.

needs prepositions and pronouns to fill their respective syntactic slots (in pp's and np's) even in the case of surface form contractions such as *deles* (*de eles*), *num* (*em um*), *pelos* (*por os*). The resolution of contractions makes syntactic structure more transparent and makes linguistic generalisation easier. Thus, it has advantages both for the disambiguation grammar (verbal valency can “see” a given preposition, nouns can “see” their article) and facilitates tasks like np-extraction,

33 PALAVRAS was the overall best-performing system in Linguatca's first and second HAREM-evaluation.

cross-language word alignment and the extraction of collocation patterns in lexicography. However, while the parser automatically splits contractions with the prepositions *de* (~52%), *em* (~37%), *a* (2.6%) and *por* (1.7%), as well the historical forms with *com* (2%) and some frequent contractions with *para* (5%), it did not cover all combinations with the latter, and obviously missed out on contractions with non-standard second parts, such as *naquea* (*em aquela*). Therefore, the remaining forms had to be expanded by the C-ORAL-preprocessing, either (a) before or (b) after PALAVRAS' tokenization step. For pre-tokenization, a regular expression-match was used, and a <contraction:...> meta-tag inserted in front of the expansion, which all were 2-part expansions, given below with their corpus frequencies:

pa (477), pro (122), co (23), pros (20), prum (18), pos (18), ca (15), pras (14), cum (9), cos (7), des (8), cos (7), pum (6), puma (5), naquea (5), cuma (5), pruma (4), dea (3), daquea (3), pas (1), naques (1), daqueas (1).

eu falei isso naquea reunião lá

eu [eu] PERS M/F 1S NOM @SUBJ>
falei [falar] <vH> V PS 1S IND VFIN @FMV
isso [isso] <dem> SPEC M S @<ACC
<contraction:naquea>

em [em] PRP @<ADVL
aquela [aquele] <dem> DET F S @>N
reunião [reunião] <occ> N F S @P<
lá [lá] ADV @N<

The most problematical case was *pra*, because this form is ambiguous - it can either be a shortened version of *para*, or a two-word contraction, *para a*, calling for contextual disambiguation.

/ só **pra** eles mesmos // (=para)
// **pra** próxima taça / (=para a)

Post-tokenization (*coral.inter*-program) was used for the contractions that were less regular and/or more difficult to match with regular expressions. These cases were drawn from C-ORAL's normalisation lexicon, and their parts were word-form numbered and marked with OALT normalization tag (cp. chapter 4), in theory allowing any number of parts:

pa despesa é bastante / né //

pa OALT pra [para] <sam-> PRP @ADVL>
a [a] <artd> <-sam> DET F S @>N
despesa [despesa] <mon> N F S @P<
é [ser] <vK> V PR 3S IND VFIN @FMV
bastante [bastante] <nh> ADJ M/F S @<SC

\$,
<slash>
né OALT não [não] ADV @ADVL>
né-2 OALT é [ser] V PR 3S IND VFIN @FMV
\$;

Note the standardization of *pa* as *pra*, not *para*, leaving it to PALAVRAS to assign and resolve the *para=a/para* ambiguity.

Portuguese, especially in spoken language, employs a special focus construction with *ser que*, which in some cases formally lends itself to a syntactic analysis with an absolute relative clause (*o que*), but more often than not, usage support a simpler, more functional analysis with *e=que*, *foi=que* or simply *é* as a focus particle, inserted before the focused constituent:

é uma cerveja que quero
--> *uma cerveja é que quero*
-> *quero é uma cerveja*

In the C-ORAL-Brasil corpus, the focus particle *é=que*, occurs 380 times, in ~ 2% of turns, but is transcribed as *que*. Since PALAVRAS would read an ordinary *que* as a conjunction or relative, rules would run into difficulties due to the absence of a subordinate clause. Therefore, the

normalization preprocessor tries to match sequences of *qu*-words and *que* (que que, quando que, quanto que, quem que, onde que), and insert the standard *e=que*, while retaining the substituted sequence in an <elision:....> meta tag:

*LEO: <beleza> // <então a gente já sabe em quem que a gente vai colocar a> culpa //

```
<LEO:>
<overlap-start>
beleza  [beleza] <am> N F S @NPHR
<overlap-stop>
$;
<overlap-start>
então  [então] <kc> ADV @ADVL>
a      [o] <artd> DET F S @>N
gente  [gente] <HH> N F S @SUBJ>
já     [já] ADV @ADVL>
sabe   [saber] <fmc> V PR 3S IND VFIN @FMV
em     [em] PRP @ADVO>
<elision:quem_que>
quem   [quem] <interr> SPEC M/F S/P @P<
é=que  [é=que] <foc> ADV @<FOC
a      [o] <artd> DET F S @>N
gente  [gente] <HH> N F S @SUBJ>
vai    [ir] V PR 3S IND VFIN @FAUX
colocar [colocar] <vH> V INF @IMV @#ICL-AUX<
a      [o] <artd> DET F S @>N
<overlap-stop>
culpa  [culpa] <am> N F S @<ACC
$;
```

3.11 Lexical and orthographic normalization

In order to assign a morphological tag string and word class hypothesis, PALAVRAS tries to recognize unknown words as either (1) affix-derivations or (2) variations of standard forms, or a combination of both. Even for written language data, (2) is an important robustness factor because of the spelling differences between European and Brazilian Portuguese (a), oi-ou and accent variation (b) etc., as well as the need to understand texts with a spelling that has been rendered obsolete by orthographic reform. Even some typos and historical forms are handled this way. The recognized standard form will be juxtaposed to the original word form as an ALT tag:

- (a) dicção ALT dição [dição] <sem-s> N F S
- (b) negócio ALT negócio [negócio] <act-d> N M S

This way, the full tag type scheme for the C-ORAL-Brasil annotation will look like this:

wordform (ALT normalization) [lemma] <secondary tags> PoS MORPHOLOGY @SYNTAX

In some rare cases, the analyzer will change an unknown, but existing word in order to match a derivational analysis, e.g. read *hemácia* as *hemacia* (*hem-ac-ia*). While leading to a lexeme error, this method will still in most cases yield a correct PoS and morphological analysis.

For the C-ORAL project, however, ordinary standardization was deemed not to be enough, first of all because certain oral word forms were transcribed in a phonetic fashion *as is*³⁴, creating in some cases unrecoverable differences from standard orthography, or the risk of ambiguity. As a side consideration, we also wanted to account for lexical gaps due to dialectal or otherwise rare forms. Therefore, two new modules were added to PALAVRAS' program chain, both with a manually maintained lexicon-file as input. The first program (*coral.inter*) handles specific or systematic standardizations and is run after preprocessing, before morphological analysis, while the second program (*postlex_pt*) is regular morphological analyzer in its own right, with its own lexicon and inflexion rules, overriding PALAVRAS' own analysis, removing the error risk created by heuristic readings.

An example for systematic normalization is the addition of first person plural -s for verbs (*comemoramo* -> *comemoramos*, *encontramo* -> *encontramos*), which *coral.inter* accomplishes using string matches and a fullform lexicon that helps to avoid false s-additions to e.g. nouns like *balsamo*, *dinamo*, *esperramo*. l-r variation (*glandão* - *grandão*) was also covered but proved to be negligible in quantitative terms.

About 700 normalizations were listed in a special lexicon file³⁵, and though the standard analyzer could have handled a certain proportion on its own in terms of word class, the lexicon treatment also allowed us to add correct base forms or even semantic classification. A very phonetic example are abbreviations (a1-3) where even plural (a2) forms and non-standard pronunciation (a3) were

34 Transcription of oral data has to strike a balance between standardization and phonetic fidelity. Too little standardization will make the corpus difficult to use and search, to do lexical frequency analysis or word order studies. Too little phonetic fidelity, on the other hand, will remove some of the very features and patterns we might want to learn from the corpus. Thus, a question like “how common is ‘-im’ as a diminutive?” or “how common is s-drop in verbal inflexion?” can obviously not be answered if full normalization is used. Therefore, only two level annotations like the one we propose, allowing both form- and category-searches at the same time, may hope to combine the best of two worlds.

35 Most of the original content for both the normalization file and add-on lexicon file was provided by one of the C-ORAL authors, Heliana Mello, followed by consistency and compatibility checks for the individual items, ensuring full tag coverage and preventing unwanted interferences with PALAVRAS' main lexicon.

covered³⁶. Other groups cover non-standard inflexion (d1-3) and derivation (c1-2). Finally, word-initial changes like a-drop (b2-4) had to be covered in order to prevent such forms from being guessed as (most likely) singular nouns.

(a1)	emedebê	MDB
(a2)	emeeles	ml
(a3)	emitivi	MTV
(b1)	envinha	vinha
(b2)	garrou	agarrou
(b3)	inda	ainda
(b4)	roz	arroz
(c1)	espim	espinhos
(c2)	ladim	ladinho
(d1)	estudemo	estudamos
(d2)	fazido	feito
(d3)	fize	fiz

While maintaining the original word form, standardized forms were added with an OALT:... prefix, and it is the standard form that annotation tags refer to:

meninim OALT menino [menino] <DERS> N M S

The standardization lexicon also covers multi-word strings (*a'=aqui -> olha=aqui, c'=ocês -> com=vocês*), which is why the tokenizer preprocessor also needs access to the file. One advantage of multi-word normalization is that the individual parts provide disambiguation context for each other, allowing, for instance, the recognition of *a'* as *olha'*, rather than the preposition or determiner reading, or the resolution of *n'* as *não* or *em* in *n'=era* and *n'=ocê*, respectively.

The second lexical add-on program, the override analyzer, is considerably more sophisticated than the normalization program, and allows both fullform and base form entries in its lexicon (*newlex_pt*). Regular inflexions of noun, adjective and verb forms will be recognized from the base form alone, but all irregular forms have to be entered separately. Like for the standardization lexicon, multi-word entries will also be visible to the preprocessor for tokenization (d1, b).

In the actual lexicon (currently 2000 entries), due to the good coverage of PALAVRAS, there are very few regular Portuguese nouns, and those there are could mostly have been recognized by PALAVRAS' derivational analysis (a1). Still, some inflected and complex forms (a2-3) may be useful to avoid the choice of a competing heuristic analysis, e.g. *caça-talentos* as plural- vs. singular-inflected. Also, the corpus contained a certain number of foreign words which are likely to be singular nouns, but may have endings that could trigger a heuristic (Portuguese) analysis as something else, e.g. *remote* (c1). Even more important is it to list foreign non-noun words such as verbs (c3), adjectives (c4) or adverbs (c5), but these entries raise two problems that would have to be resolved if the lexicon were to be used in a more general setting (i.e. for other corpora): First, foreign words would need to be specified with *all* their readings, not only the one occurring in the corpus, e.g. *shift* (c4) as both noun and verb. Second, also foreign entries would need full morphology, if they were to fully interact with their Portuguese context and CG-rules (e.g. agreement issues). This latter consideration has already been taken into account by semi-automatically adding singular male features (N M S) to noun entries without pre-entered morphology, but a similar strategy would be more difficult for verbs and adjectives due to the fact

36 PALAVRAS does handle phonetic abbreviation spelling to, but only for base forms of type (a1), and by analyzing letters as "suffixes" (<DERS>): emedebê "M" <DERS D> <DERS B> b

that English under-specifies adjective number and - in many forms - verb finity.

The largest portion of the lexicon, however, amounting to two thirds of all entries, are proper nouns (e1-3). Though these could be fairly safely recognized as such by PALAVRAS, their gender (and possibly number) is not easy to guess (e.g. *TIM* as feminine), and the addition of a semantic prototype reading (e.g. <hum>=human, <org>=organization, <Lciv>=town or state) provided valuable semantic context for CG rules, allowing, for instance, to unify the \pm HUM feature on verbs and their subjects, allowing semantics-based disambiguation of word-class or syntactic function.

- (a1) fazeção <activity> N F S
- (a2) zenes N M P # termo de jogo
- (a3) caça-talentos N M S
- (a4) superbonitinha ADJ F S
- (a5) superbem-arrumada ADJ F S
- (b) mil-oitocentos-e=vovó=gostosa NUM M/F P
- (c1) remote N M S # estrangeirismo
- (c2) completed ADJ M/F S/P # estrangeirismo
- (c3) save V # estrangeirismo
- (c4) shift N M S # estrangeirismo
- (c5) anche ADV # estrangeirismo
- (d1) tu=tu X # onomatopéia
- (d2) tuf X # onomatopéia
- (e1) Titina <hum> PROP F S
- (e2) TIM <org> PROP F S # operadora de telefonia
- (e3) Timoftol <cm-rem> PROP M S
- (f) agadê N M S # HD (harddisk)

In principle, the override lexicon module could be regarded as a general lexicon extension for PALAVRAS, since most specific adaptations, such as the afore-mentioned phonetically spelled abbreviations (f), or the female form of adjectives (a4-5), while not in standard format, do not disturb the morphological system of PALAVRAS either. On the other hand, the list contains a few entries with no regular word class, such as the onomatopoeia *tu tu* and *tuf* (d1-2), and the treatment of numerical expressions as wholes (b) is in conflict with the otherwise slightly more analytical approach used by PALAVRAS. Therefore, these word types, as well as the ambiguity potential of foreign words, should be consistency-checked before porting the lexicon to other corpora.

Originally, the C-ORAL lexicon extension was intended as a “hard override”, i.e. the idea had been to use the provided analysis *instead of* the original PALAVRAS analysis, assuming the latter would be heuristically wrong or underspecified. However, introducing a new reading, inspired by corpus inspection, because PALAVRAS did not provide the desired analysis in a particular utterance, does not, for ambiguous words, necessarily mean that PALAVRAS would have come up with the wrong analysis every time (i.e. in other contexts). And since it is more difficult for a human to come up with a full cohort of ambiguous readings than for a computer (the brain simply filters out contextually meaningless alternatives), a more cautious solution had to be chosen, where the C-ORAL lexicon *adds* to PALAVRAS' own suggestions, rather than entirely replacing them. Since this is done *before* CG disambiguation, the grammatical rules then get a chance to choose the contextually best reading. At the same time, a bias was introduced that allowed the C-ORAL lexicon (marked <newlex>) to override PALAVRAS readings marked as heuristic³⁷ more easily than those supported by the PALAVRAS core lexicon and inflexional rules. An example of such unintended ambiguity interference is the word “pô”, listed in the C-ORAL lexicon as an

37 e.g. carrying <heur> or derivation tags (<DERS>, <DERP>), or lacking a frequency tag.

interjection: Because the general conventions of the C-ORAL-Brasil transcription allowed plural inflexions of interjections, this meant that the word form “pôs” would also be tagged as an interjection, competing with the common, verbal reading³⁸.

3.12 Syntactic segmentation

While written language data provide paragraph markers, line breaks, full stops and other punctuation to deduce syntactic and informational structure, such segmentation is implicit rather than explicit in spoken language transcriptions. Thus, the necessary information to segment speech resides in prosody (i.e. rhythm, stress and intonation) as well as nonverbal signals. Depending on whether and how this information is encoded in the transcription, a parser may simply lack the segmentational information to work properly. Some speech corpora, such as the NURC corpus version described in (Bick 1998), use orthographic means to express vowel length (*'u::m'*), stress (*'esnoBAR'*) and even pauses (*'eee'*), adding further word recognition difficulties, and the need for the insertion and contextual disambiguation of pauses on the one side and true syntactic breaks on the other. In the C-ORAL corpus, on the other hand, rather than embedding prosodic information in the orthography, prosodic segmentation was marked explicitly, at transcription time, using three different segmentation strengths:

- major prosodic breaks (//), separating what functionally could be called utterances, equivalent to written language sentence separation.
- discontinuation breaks (+) between utterances
- non-terminal prosodic breaks (/), separating what could be viewed as informational units

Rather than making this information invisible to the parser by turning it into meta-tags (the strategy chosen for syntactic noise such as false starts and reiterations), we decided to replace the prosodic markers with standard punctuation, using a semicolon as the most obvious equivalent to the // terminal breaks (alternating with '...' for interruptions), and a comma for the non-terminal breaks (/). Portuguese orthography does not use obligatory commas in all places where our transcription had a slash, but inspection of annotation results showed that the extra commas helped rather than hurt. In CG-terms, a comma is a member of the BARRIER set in many context rules, separating phrase-internal material from tokens belonging to another phrase. It is therefore the global context rules that stand to profit from the introduction of prosodic punctuation marks. The positive effect is more pronounced for syntax than morphology, due to the longer dependency spans needed for the disambiguation of the former.. In the absence of explicit prosodic break markers, Bick (1998) introduced the idea of “dishesion markers” based on pauses, stress markings and hesitation interjections (eh, éh), which were tagged and disambiguated as either <break> or <pause> tags, where the former constitutes a clause or sentence break, while the latter does not, and is allowed inside clauses and even phrases. Dishesion markers can be inserted near prosodic annotation features, but can also be mapped on regular words, using ordinary CG context rules to define syntactic edges, such as conjunctions for clauses, prepositions for pp's and articles and determiners for noun phrases. For the C-ORAL-Brasil corpus this technique was implemented exploiting the corpus' explicit prosodic markers. The // “major break” was substituted with a semicolon, while the / “soft break” was re-tagged as a comma with two potential readings, <break> and <pause>, where only the former represents a syntactic break, while the latter is allowed inside phrases and between verb and complement. CG-rules were written to distinguish between these two readings, and the comma replaced with a meta-tag for the <pause> cases - making it invisible to ordinary CG disambiguation rules. Contextually disambiguating the function of prosodic breaks allowed us to strike a balance between simply ignoring such markup on the one hand, and syntactic over-

38 and possibly the adverb *pois*, which however was not listed as *pôs* in the standardization lexicon.

segmentation on the other. If the original parser rules were to work optimally, they needed a comma marker that was as close to a standard written language comma as possible.

The following are clear-text versions of some of the CG rules used for this disambiguation task:

1. a prosodic /-marker is treated as <break> if it occurs before the first word of an np, or before a pronoun in the nominative, followed by a finite verb to the right (i.e. clause-initially)
2. between a noun or a nominative pronoun to the left, and a finite verb to the right, a prosodic /-marker is treated as <pause> (subject - verb case)
3. prosodic /-markers between a noun and another np are treated as <break> (appositions)
4. prosodic /-markers between potential np-parts are treated as <pause> if there is gender-number agreement between the np part candidates (e.g. DET-ADJ-N)
5. between a noun and an adjective agreeing in gender and number, a prosodic /-marker is treated as <pause> (i.e. N ADJ)
6. between a transitive verb and a left np edge, a prosodic /-marker is treated as <pause>
7. between an auxiliary and its main verb, a /-marker is treated as <pause>
8. if a single word is surrounded by prosodic /-markers, these are treated as <pause>
9. if a prosodic /-marker is preceded by a conjunction or relative, it is treated as <pause>
10. if a prosodic /-marker is preceded by a preposition, it is treated as <pause> (i.e. pp-internal)
11. between an intensifier and an attribute, '/' is treated as <pause> (i.e. adjective phrase-internal)
12. if a prosodic /-marker is followed by certain, typically postnominal prepositions (de, em, com, sem) in certain contexts, it is treated as <pause>

3.13 Evaluation

In order to evaluate the modified parser on our data, one transcription file (bfamd115) was chosen at random, automatically analyzed and hand-corrected. We then used the Constraint Grammar evaluation tool `eval_cg` to compare the raw analysis file with the revised version. In an ordinary CG setup, meta-markup and punctuation would align 100%, but in our case, matters were complicated by the fact that “commas” had been disambiguated as either break or pause, and in the latter case replaced with a meta-tag. On the one hand, this caused alignment problems for the evaluator, on the other hand, differences had to be identified and counted as recall errors. Other mismatches, caused by faulty splitting or non-splitting of ambiguous MWE's, were also counted as recall errors, e.g in the case of “*primeiro=que*” (conjunction vs. adjective/numeral + relative). Including “punctuation” tokens, the file contained 1895 word tokens.

	Recall	Precision	F-Score
Syntactic function	95,3	94,9	95
PoS / Word class	98,5	98,7	98,6
Morphology	98,4	98,6	98,5
Base form	98,6	99,4	99

Table 3: Parser performance

It can be seen from these figures that the easiest task was lemmatization (base forms), while syntactic function was the most difficult. The difference between recall and precision for syntax is a measure of remaining ambiguous tags. For word class and morphology, only one reading was allowed, so the precision-recall differences are entirely due to differences in matching differences between break markers (commas).

In order to judge the effectiveness of using prosodic break markers as punctuation, we also compared the standard run (with pause/break disambiguation) with a no-break run (/marks ignored), a no-sentence run (both /, + and // ignored), and an all-break run (all /marks turned into commas, *without* disambiguation). Since the gold file did have disambiguated commas, the evaluator was run in match-only mode, comparing tags only for matching tokens. Therefore, figures in the table below can only be compared with each other, and not with the original test run³⁹.

	no-sentence	no-break	all-break	pause / break
Syntactic function	86.2 (R: 86.5, P: 86.1)	90.7 (R: 91.0, P: 90.6)	93.7 (R: 93.3, P: 93.6)	95.0 (R:95.3, P: 94.8)
PoS / Word class	98,3	98,8	99,3	99,4
Morphology	98,1	98,6	99	98,7
Base form	99	99,1	99,4	99,4

Table 4: Influence of syntactic break markers

Clearly, exploiting prosodic break markers did improve performance at all levels. However, the effect was much more marked for syntax than for part of speech, lemmatization and morphology, reflecting the wider contextual scope of syntactic tags and the ensuing greater need for precise and correct segmentation⁴⁰. Interestingly, while syntactic performance can be further increased by pause/break disambiguation, this is not obvious for the more local tag categories. Thus, for inflexion tags (morphology), all-break performance was *higher* than for the pause/break run, and only for part of speech a slight improvement was observed.

Given the large size of the tag set and the relatively low error rate, it is difficult to establish statistically significant error tendencies, without a manual revision of large parts of the corpus. For what it's worth, a confusion matrix is presented below. Left numbers concern the best case scenario, right numbers are from a raw run without the syntactic use of prosodic break markers.

correct	V	N	ADJ	ADV	DET	PERS	SPEC	PRP	NUM	KS	IN
error											
V	###	0-1		2-3							
N	1-3	###			0-1						
ADJ			###	0-1							
ADV				###	2-2						0-4
DET				2-3	###		1				
PERS					1-1	###				2-2	
SPEC				0-1			###			2-1	
PRP								###			
NUM					0-1				###		
KS						0-2	0-1			###	
IN											###

Table 5: Confusion matrix for PoS

³⁹ Also, this experiment was done at a later stage, where some improvements in the general grammar had been made, making a direct comparison impossible.

⁴⁰ It can be concluded that the positive effect of such segmentation on uniqueness principle rules and NOT-rules (which profit from more fine-grained segmentation) outweighs the potential comma-blocking of *positive* rules looking for long-distance syntactic relations.

The matrix shows that the most serious error, verbo-nominal confusion, can be considerably reduced through the use of prosodic break markers, and the same is true for interjections and the disambiguation of *que* and *se* as a conjuncton or pronoun, respectively.

It should be noted, that in dubious cases, the error evaluation was done *in dubio pro reo*, accepting, for instance, the parser's morphological word class definition as correct, even where an alternaive classification might be preferred, if the syntactic tag would allow filtering into the latter. Thus, participle readings for *errado*, *separado* or *recheado* were accepted as equivalent to adjective readings, if combined with ad-nominal or predicative function readings (@N<, @PRED ...) rather than functionally verbal readings (@ICL-AUX<).

During development, a certain amount of "error artefacts" were observed, resulting from the introduction of new ambiguity from the add-on lexicon, which for instance listed *é* as a noun and *seu* a form of "senhor". These cases were, however, largely remedied in the final version of the corpus.

For syntactic function, more errors could be observed than for PoS, but very few confusion pairs had more than one error within the evaluation chunk, indicating a large and somewhat unsystematic type spread⁴¹. The most frequent cases are listed below. '*' marks function-only errors (i.e. with correct dependency attachment), '+' marks valency-only errors (i.e. same function, but argument-adjunct confusion).

Erroneous tag	correct tag	cases in chunk	explanation
@ADVL>	@<ADVL	5	adverbial attachment left/right
@SUBJ>	@<ACC	4	subject-object ambiguity between verbs
@<SC*	@<ACC	4	predicative vs. object
@<SUBJ*	@<ACC	3	subject-object ambiguity right of a verb
@<ADVL*+	@<ADVO	3	free versus object-bound adverbial
@<ADVL	@A<	3	adverbial adjunct or adverbial attribute
@VOK	@<SC	2	vokative or predicative
@VOK	@APP	2	vokative or apposition
@SUBJ>*	@ADVL>	2	subject or noun (time?) adverbial
@PRED>*	@SUBJ>	2	subject or free predicative
@PRED>	@NPHR	2	free predicative or isolated noun phrase
@<PIV*+	@<ADVL	2	pp as object or adverbial
@FMV	@FOC>	2	'é' as verb or focus marker
@<ADVL	@>N	2	free adverbial or prenominal
@<ADVL*+	@<ADVS	2	free versus subject-bound adverbial

Table 6: Confusion matrix for syntactic function

As can be seen, adverbial function tags are relatively problematic, though the errors mostly stem from the ±valency distinction (ADVL vs. ADVS, ADVO, PIV), dependency direction (<, >) and head type (ADVL vs. A<), while adverbiality as such is rarely a problem (ADVL vs. SUBJ, probably time nouns). Another error group with mostly group-internal confusion concerns clause-external material not part of ordinary verb frames (VOK, PRED, NPHR, APP). Among ordinary, verb-linked clause functions, subject-object confusion is a common problem between verbs, and

⁴¹ As a consequence, improvements to the parsers would have to address many individual error types rather than a few obvious and frequent cases.

predicative-object confusion to the right of a verb. While the former represents a dependency error at the same time, the latter does not constitute a dependency attachment error.

PALAVRAS also uses some 200 semantic prototype tags for nouns, not listed here, as well as valency tags for verbs, nouns and adjectives.

REFERENCES

- Austin, L.J., 1962. *How to do things with words*. Oxford: Oxford University Press.
- Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E. 1999. *The Longman grammar of spoken and written English*. London: Longman.
- Bick, Eckhard. 2000. The Parsing System Palavras - Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework, Aarhus: Aarhus University Press
- Bick, Eckhard. 1998. Tagging Speech Data - Constraint Grammar Analysis of Spoken Portuguese, in: Proceedings of the 17th Scandinavian Conference of Linguistics (Odense 1998)
- Bick, Eckhard & Marcelo Módolo. 2005. Letters and Editorials: A grammatically annotated corpus of 19th century Brazilian Portuguese. In: Claus Pusch & Johannes Kabatek & Wolfgang Raible (eds.) *Romance Corpus Linguistics II: Corpora and Historical Linguistics (Proceedings of the 2nd Freiburg Workshop on Romance Corpus stics, Sept. 2003)*. pp. 271-280. Tübingen: Gunther Narr Verlag.
- Brill, Eric. 1992. A simple rule-based part of speech tagger. In: Proceedings of the workshop on Speech and Natural Language. HLT '91, Morristown, NJ, USA: Association for Computational Linguistics, pp.112-116
- Castilho, Ataliba de (ed.), 1993. Gramática do Português Falado, vol.3, Campinas: Editora da Unicamp.
- CHAT <http://chilides.psy.cmu.edu/manuals/CHAT.pdf>
- Cresti, E., 1994. Information and intonational patterning in Italian. In *Accent, intonation, et modèles phonologiques*, B. Ferguson, H. Gezundhajt, Ph. Martin (eds.), 99-140. Toronto: Editions Mélodie.
- Cresti, E. 2000. *Corpus di italiano parlato*, voll. I-II, CD-Rom. Firenze: Accademia della Crusca.
- Cresti, E.; Moneglia, M. (Eds.) (2005). *C-ORAL-ROM: Integrated Reference Corpora For Spoken Romance Languages*. Amsterdam: John Benjamins.
- Crystal, D. 1975. *The English tone of voice*. London: Edward Arnold.
- De Mauro, T. et alii 1993. *Lessico di frequenza dell'italiano parlato*. Milano: ETAS libri.
- 't Hart J., Collier R., Cohen A. 1990. *A perceptual study on intonation. An experimental approach to speech melody*. Cambridge: Cambridge University Press.
- Houaiss, A. 1990. Novo Dicionário Houaiss in the CD-ROM version. Rio de Janeiro: Editora Objetiva.
- Johannessen, Janne Bondi, Joel Priestley, Kristin Hagen, Tor Anders Åfarli, and Øystein Alexander Vangsnes. 2009. The Nordic Dialect Corpus - an Advanced Research Tool. In Jokinen, Kristiina and Eckhard Bick (eds.): Proceedings of the 17th Nordic Conference of Computational Linguistics NODALIDA 2009. NEALT Proceedings Series Volume 4
- Karcevsky, S. 1931. "Sur la phonologie de la phrase". In *Travaux du Cercle linguistique de Prague* IV, 188-228.
- Karlsson, Fred & Voutilainen, Atro & Heikkilä, Juka & Anttila, Arto. 1995. Constraint Grammar, A Language-Independent System for Parsing Unrestricted Text. Berlin: Mouton de Gruyter.
- Maamouri, Mohamed et al. 2010. From Speech to Trees: Applying Treebank Annotation to Arabic Broadcast News. In: Proceedings of LREC 2010, Valletta, Malta, May 2010.
- MacWhinney, B. 1994. *The CHILDES project: tools for analyzing talk*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.
- Martin. P. 2011. WinPitch Pro <www.winpitch.com>.

- Mello, H.; Raso, T. (2009). Para a transcrição da fala espontânea: o caso do C-ORAL-BRASIL. *Revista Portuguesa de Humanidades*, v. 13, n. 1, p. 153-178. Disponível em: <<http://www.ufjf.br/revistaveredas/files/2009/11/ARTIGO-Tommaso-Raso-e-Heliana-Mello.pdf>>.
- Miller, J. and Weinert, R. 1998. *Spontaneous Spoken language*. Oxford: Clarendon Press.
- Moreno, A. & J.M. Guirão. 2003. "Tagging a spontaneous speech corpus of Spanish". In: *Proceedings of the International Conference on Recent Advances in Natural Language Processing*. Borovets, Bulgaria, 2003. p. 292-296.
- Müürisep, Kaili and Uibo, Heli (2006). "Shallow Parsing of Spoken Estonian Using Constraint Grammar". In: P.J.Henriksen & P.R.Skadhauge, *Proceedings of NODALIDA-2005 special session on treebanking*. Copenhagen Studies in Language #33/2006
- Moneglia, M., A. Panunzi, E. Picchi, 2004, *Using PiTagger for Lemmatization and PoS Tagging of a Spontaneous Speech Corpus : C-Oral-Rom Italian*. In M.T. Lino et al. (eds.), *Proceedings of the 4th LREC Conference*, vol. 2, ELRA, Paris, pp. 563-566.
- Quirk, R., Greenbaum, S., Leech, G., Svartvik, J. 1985. *A comprehensive grammar of the English language*. Longman: London.
- Raso, Tommaso & Heliana Mello. 2010. The C-ORAL BRASIL corpus. In: Massimo Moneglia & Alessandro Panunzi (eds): *Bootstrapping Information from Corpora in a Cross-Linguistic Perspective*. Università degli studi di Firenze, Biblioteca Digitale.
- Schmid, Helmut. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of the International Conference on New Methods in Language Processing 1994*. pp. 44-49.

Appendixes

Appendix 1- Typical examples of prosodic breaks types in Brazilian Portuguese ⁴²

Examples of strings with typical non terminal breaks, generic terminal breaks and interruption (highlighted in the examples).

Example of terminal break (bpubmn01)

*GUI: é a terceira //\$ vão lá //\$ foi //\$ isso //\$

Examples of strings with typical interruption (bfamcv01)

*LUI: <eu acho não> //\$

*LEO: <com certeza> //\$

*LUI: <com certeza es nũ vão participar / uai> //\$

*LEO: <eles são piores do que o> Durepox //\$

*EVN: é / pois <é> //\$

*LUI: <agora> manda uma barrinha <minha> //\$

*EVN: <porque o Durepox> / pelo menos jogava bola //\$

*GIL: não / e o Durepox / eu vou +\$ tinha um cara //\$ era &aque [2] era &aque [2] era aquele cara <lá / que era muito> / <muito> / muito <palha> //\$

*EVN: <era aquele cara / é> //\$

*PAU: ah /=EXP= porque senão //=COM=\$ aqui o' //=COM=\$ aí por exemplo +=TOP=\$

*ROG: aqui já tá dando [4] aqui já tá dando a altura //\$

Examples of strings with typical retracting (bpubdl01)

*PAU: ah /=EXP= porque senão //=COM=\$ aqui o' //=COM=\$ aí por exemplo +=TOP=\$

*ROG: aqui já tá dando [4] aqui já tá dando a altura //\$

Examples of strings with typical non-terminal break (bfamdl32)

*BAL: quando sai / nũ é stop //\$

⁴² The Multimedia files corresponding to the examples are available in the directory *Appendix* of the DVD.

Appendix 2 - Tagsets used for PoS tagging in brazilian portuguese. Detailed tables and comparison table

1. Verbs

V Verb

person/number

1S first person singular
2S second person singular
3S third person singular
1P first person plural
2P second person plural
3P third person plural
0/1/2/3S zero-morpheme infinitive

mood

IND indicative
SUBJ subjunctive
COND conditional
IMP imperative

tense

PR present
IMPF past (imperfecto)
FUT future (simple)
PS past (perfeito simples)

non-finite forms

INF infinitive (+ number/gender)
GER gerund
PCP participle (+ number/gender)

2. Nominals and pronominals:

N Noun
PROP Proper noun
ADJ Adjective
DET Determiner
NUM Numeral

gender

M masculine
F feminine
M/F invariable/underspecified

number

S singular
P plural
S/P invariable/underspecified

secondary nominal classes

<KOMP> comparative
<SUP> superlative
<NUM-ord> ordinal
<card> cardinal

PERS Personal pronoun

person/number

1S	first person singular
2S	second person singular
3S	third person singular
1P	first person plural
2P	second person plural
3P	third person plural

gender

M	masculine
F	feminine
M/F	invariable/underspecified

case

NOM	nominative
ACC	accusative
DAT	dative
PIV	prepositive
NOM/PIV	underspecified
ACC/DAT	underspecified

INDP Independent (non-inflecting)

Secondary pronoun classes

<arti> DET	indefinite article
<artd> DET	definite article
<dem> DET, INDP	demonstrative
<poss...> DET	possessive
<quant> DET	quantifier
<rel> DET, INDP	relative
<interr> DET, INDP	interrogative
<refl> PERS	reflexive personal pronoun
<si> DET	reflexive possessive

3. Non-inflecting word classes

ADV Adverb

<rel>	relative
<interr>	interrogative
<ks>	similar to subordinating conjunction
<kc>	similar to coordinating conjunction
<foc>	focus marker

PRP Preposition

KS Subordinating conjunction

KC Coordinating conjunction

IN Interjection

4. Syntactic tags

@SUBJ> @<SUBJ subject

@ACC> @<ACC accusative (direct) object

@DAT> @<DAT dative object (only pronominal)

@PIV> @<PIV prepositional (indirect) object

@ADVS> / @SA> @<ADVS / @<SA adverbial predicative (place, time, duration, quantity), equivalent to nominal @SC

@ADVO> / @OA> @<ADVO / @<OA adverbial object predicative, equivalent to nominal @OC

@SC> @<SC subject predicative

@OC> @<OC object predicative

@ADVL> @<ADVL adverbial
 @PASS> @<PASS agent of passive
 @ADVL 'free' adverbial phrase (in non-sentence expression)
 @NPHR 'free' noun phrase (in non-sentence expression without verbs)
 @VOK 'vocative' (e.g. 'free' addressing proper noun in direct speech)
 @>N prenominal adjunct
 @N< postnominal adjunct
 @N<PRED postnominal (in-group predicative)
 @APP identifying apposition
 @>A prepositioned adverbial adjunct
 @A< postpositioned adverbial adjunct
 @PRED> 'forward' free predicative
 @<PRED 'backward' free predicative
 @P< argument of preposition
 @S< sentence-related "apposition"
 @FAUX finite auxiliary (cp. @#ICL-AUX<)
 @FMV finite main verb
 @IAUX infinite auxiliary (cp. @#ICL-AUX<)
 @IMV infinite main verb
 @PRT-AUX< verb chain particle (preposition or "que" after auxiliary)
 @CO coordinating conjunction
 @SUB subordinating conjunction
 @KOMP< argument of comparative (e.g. "do que" referring to *melhor*)
 @COM direct comparator without comparative
 @PRD role predicator (e.g. "work *as*")
 @FOC> @<FOC focus marker ("gosta *é* de peixe.")
 @TOP topic constituent ("Esse *negócio*, não gosto dele.")
 @#FS- finite subclause (combines with clausal role and intra-clausal word tag, e.g. @#FS-<ACC @SUB for "não acredito *que* seja verdade")
 @#ICL- infinite subclause (combines with clausal role and intra-clausal word tag, e.g. @#ICL-SUBJ> @IMV in "*consertar* um relógio não é fácil")
 @#ICL-AUX< argument verb in verb chain, refers to preceding auxiliary
 @#AS- averbal (i.e. verb-less) subclause (combines with clausal role and intra-clausal word tag, e.g. @#AS-<ADVL @ADVL> in "ajudou *onde* possível")
 @AS< argument of complementiser in averbal subclause

5. Some secondary tags

<*> upper case
 <*1> left quote
 <*2> right quote
 <sam-> first part of contraction
 <-sam> second part of contraction
 <parkc-1> first part in *ou...ou*
 <parkc-2> second part in *ou...ou*
 <pp> complex adverb (prepositional phrase)
 <hyfen> hyphenated word
 <newlex> from add-on lexicon

Synopsis tag sets

The tags for Brazilian Portuguese below are the ones which are comparable to those present in the C-ORAL-ROM languages. However, the complete tag set for Brazilian Portuguese exhibits a sophisticated set covering both morphosyntactic and semantic categories.

Table 5a. Synopsis of the PoS tag sets

Tag-Set Projection	French	Italian	Portuguese	Spanish	Brazilian Portuguese
--------------------	--------	---------	------------	---------	----------------------

Nouns	NOM	S	N	N	N
verbs	VER	V	V	V	V
adjectives	ADJ:QUA	A	ADJ	ADJ	ADJ
adverbs	ADV	B	ADV	ADV	ADV
prepositions	PRE	E	PREP	PREP	PRP
conjunctions	CON	C	CONJ	C	KS,KC
interjections	INT	I	INT	INT	IN
discourse markers			MD	MD	
emphatic			ENF		

Table 5b. Synopsis of the PoS tag sets

Tag-Set Projection	French	Italian	Portuguese	Spanish	Brazilian Portuguese
articles	DET:DEF (definite determiner)	R (articles)	ART (articles)	DETd (definite determiner)	DET, <arti>, <artd>
demonstrative determiners	DET:DEM	DIM	DEM	DETdem	<dem> DET, INDP
demonstrative pronouns	PRO:DEM				
possessive determiners	DET:POS	POS	POS	DETposs	<poss...> DET
possessive pronouns	PRO:POS				
personal pronouns	PRO:PER	PER	PES	PPER	PERS
clitic			CL		<refl> PERS
rel-int-excl determiners	DET:INT				<interr>, <rel> DET, INDP
rel-int-excl pronouns	PRO:RIN	REL	REL	PR	
indefinite determiners	DET:IND	IND	IND		
indefinite pronouns	PRO:IND				
numbers (cardinals)	NUM	N	NUMc	Q	<card>
numerals (ordinals)	ADJ:ORD	NA	NUMo	(quantifiers)	<NUM-ord>

Table 6. Synopsis of the morpho-syntactic encodings for verbs

	French	Italian	Portuguese	Spanish	Brazilian Portuguese
MOOD	indicative	indicative	indicative	indicative	IND
	subjunctive	subjunctive	subjunctive	subjunctive	SUBJ
	conditional	conditional	conditional	conditional	COND
	imperative	imperative	imperative	imperative	IMP
	infinitive	infinitive	infinitive	infinitive	INF
	participle	participle	participle	participle	PCP
		gerund	gerundive adjectival past participle	gerund	GER
TENSE	present	present	present	present	PR
	past	past	past	past	PS
	imperfect	imperfect	imperfect	imperfect	IMPF
	future	future	pluperfect future	future	FUT
PERSON		first		first	1
		second		second	2
		third		third	3
NUMBER		singular		singular	S
		plural		plural	P
GENDER (only for participles)		masculine		masculine	M
		feminine		feminine	F
		common			
VERB TYPE		main	main	main	
		non-main	auxiliary	auxiliary	

Table 7 a. Synopsis of the non-standard tag sets

	French	Italian	Portuguese	Spanish	Brazilian Portuguese
extralinguistic	-	XLG	EL	<nl>	
paralinguistic support & fillers	-	PLG	PL	<sup>	
fragments	-		FRAG	-	

Table 7b. Synopsis of the non-standard tag sets

	French	Italian	Portuguese	Spanish
foreign words	XXX:ETR	(Pos) + K	ESTR	-
new formations	-	(Pos) + Z	-	-
acquisition forms	-	ACQ	-	-
onomatopoeia	-	ONO	-	-
meaningless forms	-		-	<nc>
euphonic particle	XXX:EUP		-	-
non understandable words	-	X	Pimp	-

Appendix 3 Orthographic transcription conventions in Brazilian Portuguese

Brazilian transcription conventions

The general orthographical norms

As a general rule, the team transcribed the entire corpus according to the official orthography.

Paralinguistic Sounds

hhh (coughing, laugh, and other not specified sounds)

nts (interjection)

psiu (interjection)

Alphabetic letter and numbers

á (name of alphabetic letter “a”; noun)

á (in “primeiro á” (school class), “papel á quatro”, “hepatite á”; adjective)

agá (name of alphabetic letter “h”; noun)

bê (name of alphabetic letter “b”; noun)

bê (in “hepatite bê”, “complexo bê”, “ônibus três bê”; adjective)

cá (name of alphabetic letter “k”, in “quarenta-e-oito cá”; noun)

cê (name of alphabetic letter “c”; noun)

cê (in “hepatite cê”; adjective)

cento-e (numeral)

cento-e-cinco (numeral)

cento-e-cinqüenta (numeral)

cento-e-dez (numeral)

cento-e-noventa (numeral)

cento-e-noventa-e-dois (numeral)

cento-e-noventa-e-três (numeral)

cento-e poucas (numeral + pronoun)

cento-e-quarenta (numeral)

cento-e-quarenta-e-sete (numeral)

cento-e-sessenta-e tantos mil (numeral + pronoun + numeral)

cento-e-setenta (numeral)

cento-e-setenta-e poucos mil (numeral + pronoun + numeral)

cento-e-trinta (numeral)

cento-e-vinte (numeral)

cento-e-vinte-e-cinco (numeral)

cento-e-vinte-e-nove-mil (numeral)

cinco e vinte-e-oito (numeral)

cinco-mil (numeral)

cinco-mil-e-seiscentos (numeral)

cinqüenta-e (numeral)

cinqüenta-e-cinco (numeral)

cinqüenta-e-nove (numeral)

cinqüenta-e-nove e noventa (numeral)

cinqüenta-e-quatro (numeral)

cinqüenta-e-sete (numeral)

cinqüenta-e-um (numeral)

cinqüenta-mil (numeral)

dábliu (name of alphabetic letter “w”; noun)

dê (name of alphabetic letter “d”; noun)
 dez e oitenta-e-sete (numeral)
 dez-mil (numeral)
 dois e cinqüenta-e-oito (numeral)
 dois e noventa-e-cinco (numeral)
 dois e noventa-e-oito (numeral)
 dois e oitenta-e-oito (numeral)
 dois e trinta-e-cinco (numeral)
 dois e trinta-e-oito (numeral)
 dois-mil (numeral)
 dois-mil-e-cinco (numeral)
 dois-mil-e-dez (numeral)
 dois-mil-e-nove (numeral)
 dois-mil-e-novecentos (numeral)
 dois-mil-e-oito (numeral)
 dois-mil-e-quatro (numeral)
 dois-mil-e-quinientos (numeral)
 dois-mil-e-seis (numeral)
 dois-mil-e-sete (numeral)
 dois-milhões-e (numeral)
 duas-mil (numeral)
 duzentas-e-cinqüenta (numeral)
 duzentos-e-quarenta (numeral)
 duzentos-e-setenta-e-sete (numeral)
 duzentos-e-trinta (numeral)
 duzentos-e-trinta-e-nove (numeral)
 duzentos-e-vinte (numeral)
 duzentos-mil (numeral)
 é (name of alphabetic letter “e”; noun)
 e-cinco (numeral)
 efe (name of alphabetic letter “f”; noun)
 ele (name of alphabetic letter “l”; noun)
 eme(s) (name of alphabetic letter “m”; noun)
 eme (in “eu vou ficar com a eme” (meaning “tamanho médio de roupa”); adjective)
 ene (name of alphabetic letter “n”; noun)
 ene (in “ene coisas” (meaning indeterminate quantity); adjective)
 ene ás (means an indeterminate quantity of “A”; adjective)
 ene és (means an indeterminate quantity of “E”; adjective)
 e-oito (numeral)
 erre (name of alphabetic letter “r”; noun)
 esse (name of alphabetic letter “s”; noun)
 e-três e sessenta (numeral)
 e vinte-e-oito (numeral)
 fê (name of alphabetic letter “f”; noun)
 gê (name of alphabetic letter “g”; noun)
 gê (meaning big size of clothes; adjective)
 ípsilon (name of alphabetic letter “y”)
 lê (name of alphabetic letter “l”; noun)
 mê (name of alphabetic letter “m”; noun)
 mil-e-duzentos (numeral)
 mil-e-quinientos (numeral)
 mil-novecentos-e-cinqüenta-e-oito (numeral)

mil-novecentos-e lá vai bolinha pa frente (numeral + adverb + verb + noun + preposition + noun)
 mil-novecentos-e nada (numeral + pronoun)
 mil-novecentos-e-sessenta (numeral)
 mil-novecentos-e-sessenta-e (numeral)
 mil-novecentos-e-sessenta-e-dois (numeral)
 mil-novecentos-e-sessenta-e-seis (numeral)
 mil-novecentos-e-setenta-e-oito (numeral)
 mil-novecentos-e-setenta-e-sete (numeral)
 mil-novecentos-e-vinte (numeral)
 mil-oitocentos-e vovó gostosa (numeral + noun + adjective)
 nê (nome da letra “n”; noun)
 nove-mil (numeral)
 noventa-e (numeral)
 noventa-e-cinco (numeral)
 noventa-e-dois (numeral)
 noventa-e-nove (numeral)
 noventa-e-oito (numeral)
 noventa-e-quatro (numeral)
 noventa-e-seis (numeral)
 noventa-e-sete (numeral)
 noventa-e-um (numeral)
 ó (name of alphabetic letter “o”; noun)
 oitenta-e (numeral)
 oitenta-e-duas (numeral)
 oitenta-e-cinco (numeral)
 oitenta-e-nove (numeral)
 oitenta-e-oito (numeral)
 oitenta-e-quatro (numeral)
 oitenta-e-seis (numeral)
 oitenta-e-sete (numeral)
 oitenta-e-três (numeral)
 oitenta-e-um (numeral)
 oito-mil (numeral)
 onze-mil (numeral)
 pê (name of alphabetic letter “p”; noun)
 pê (meaning small size of clothes; adjective)
 quarenta-e (numeral)
 quarenta-e-cinco (numeral)
 quarenta-e-cinco-mil (numeral)
 quarenta-e-nove (numeral)
 quarenta-e-nove e noventa (numeral)
 quarenta-e-oito (numeral)
 quarenta-e pouco (numeral + pronoun)
 quarenta-e-quatro (numeral)
 quarenta-e-seis (numeral)
 quarenta-e tantos (numeral + pronoun)
 quarenta-e-três (numeral)
 quarenta-e-três e sessenta (numeral)
 quarenta-e-um (numeral)
 quatorze e oitenta-e-nove (numeral)
 quatrocentos-e-oitenta-e-cinco (numeral)
 quatrocentos-e-vinte (numeral)

quatro e noventa-e-oito (numeral)
 quatro e trinta-e-dois (numeral)
 quatro-mil (numeral)
 quinhentas-e-cinqüenta (numeral)
 sê (name of alphabetic letter “s”)
 seis e quarenta-e-oito (numeral)
 seis-mil (numeral)
 sessenta-e-cinco (numeral)
 sessenta-e-dois (numeral)
 sessenta-e-dois e oitenta-e-um (numeral)
 sessenta-e nũ sei quantos (in “um ano e sessenta-e nũ sei quantos meses”; numeral + adverb + verb + adjective)
 sessenta-e-nove (numeral)
 sessenta-e-nove e noventa (numeral)
 sessenta-e-oito (numeral)
 sessenta-e-quatro (numeral)
 sessenta-e-três (numeral)
 sessenta-mil (numeral)
 setecentos-e-quatro (numeral)
 setenta-e (numeral)
 setenta-e-cinco (numeral)
 setenta-e-dois (numeral)
 setenta-e-nove (numeral)
 setenta-e-oito (numeral)
 setenta-e poucos (numeral + pronoun)
 setenta-e-seis (numeral)
 setenta-e-sete (numeral)
 setenta-e-três (numeral)
 setenta-e-um (numeral)
 tê (name of alphabetic letter “t”; noun)
 três e quarenta-e-oito (numeral)
 três e vinte-e-oito (numeral)
 três-mil (numeral)
 três-mil-e (numeral)
 três-mil-e-duzentos (numeral)
 três-mil-e-quinhentos (numeral)
 três-milhões (numeral)
 trezentos-e-cinqüenta (numeral)
 trezentos-e-dois (numeral)
 trezentos-e-três (numeral)
 trezentos-e-um (numeral)
 trinta-e (numeral)
 trinta-e-cinco (numeral)
 trinta-e-dois (numeral)
 trinta-e-dois e sessenta-e-dois (numeral)
 trinta-e-nove (numeral)
 trinta-e-nove e noventa (numeral)
 trinta-e-oito (numeral)
 trinta-e pouco (numeral + pronoun)
 trinta-e poucos (numeral + pronoun)
 trinta-e-seis (numeral)
 trinta-e-sete (numeral)

trinta-e tantas (numeral + pronoun)
 trinta-e-três (numeral)
 trinta-e-três e sessenta (numeral)
 trinta-e-três e vinte (numeral)
 trinta-e-um (numeral)
 um e noventa-e-cinco (numeral)
 um e noventa-e-nove (numeral)
 um e noventa-e-oito (numeral)
 um e oitenta-e-cinco (numeral)
 um e setenta-e-cinco (numeral)
 um e setenta-e-nove (numeral)
 um e vinte-e-oito (numeral)
 um-milhão (numeral)
 um-milhão-e-meio (numeral)
 um-milhão-e-setecentos-mil (numeral)
 vinte-e (numeral)
 vinte-e-cinco (numeral)
 vinte-e-dois (numeral)
 vinte-e-nove (numeral)
 vinte-e-nove e noventa (numeral)
 vinte-e-oito (numeral)
 vinte-e-quatro (numeral)
 vinte-e-seis (numeral)
 vinte-e-sete (numeral)
 vinte-e-sete-milhões (numeral)
 vinte-e-um (numeral)
 vinte-mil (numeral)
 vinte-milhões (numeral)
 xis (name of alphabetic letter “x”; noun)
 xis (in “nós somos a xis e a ípsilon”; noun)
 xis (in “raio X”; proper noun)
 xis (meaning “anyone” (“no lugar xis”) and indeterminated quantity (“xis coisas”); adjective)
 zero ponto oitenta-e-cinco (numeral)
 zero ponto vinte-e-três (numeral)

Initials and acronyms

abecê (in “curva ABC de produto”; proper noun)
 agabeesse (= HBs; proper noun)
 agabeesseagê (= HBsAg; proper noun)
 agadê (= HD; noun)
 agaé (= he; noun)
 agaivê (= HIV; proper noun)
 agateelevê (= HTLV; proper noun)
 ANVISA (proper noun)
 aveí (= AVI; proper noun)
 beagá (= BH; proper noun)
 beagá Shopping (= BH Shopping; proper noun)
 beerre (= BR; national highway; noun)
 CAT (proper noun)
 ceagaeme (= CHM; noun)
 cedê(s) (= CD(s); noun)

CEFET (proper noun)
 CEFET emegê (= CEFET-MG; proper noun)
 CEFET Minas (proper noun)
 CEFET Ouro Preto (proper noun)
 CEMIG (proper noun)
 CENEX (proper noun)
 cepecê (= CPC; proper noun)
 ceteí (= CTI; proper noun)
 deá (= DA; proper noun)
 deeleele (= DLL; proper noun)
 deeneá (= DNA; proper noun)
 deerrei (= DRI; proper noun)
 deessetês (= DSTs, “doenças sexualmente transmissíveis”; noun)
 dejota (= DJworking place name; proper noun)
 devedê (= DVD; noun)
 didjei (= DJ; noun)
 didjeis (= DJs; noun)
 DPVAT (proper noun)
 EDUCONLE (proper noun)
 efeele dois (= FLV 2 wrong pronunciation; proper noun + numeral)
 efeelevê (= FLV; proper noun)
 efeeme (= FM; name of a rádio; proper noun)
 EJEJ (proper noun)
 EMBRATEL (proper noun)
 emebê (= MB, in “eu coloquei emebê pra ele” (evaluation like “muito bom”); noun)
 emecá (= MK; noun)
 emeci Escher (= M.C. Escher; proper noun)
 emeci Hammer (= MC Hammer; proper noun)
 emeci Ice (= MC Ice; proper noun)
 emedebê (= MDB; proper noun) (the correct form is “iemedebê”, = IMDb)
 emeele(s) (= ml; noun)
 emeesseene (= MSN; proper noun)
 emepégê (= MPG; proper noun)
 emepegue (= MPEG; proper noun)
 emepegue dois (= MPEG-2; proper noun + numeral)
 emepegue um (= MPEG-1; proper noun + numeral)
 emepê três (= MP3; proper noun + numeral)
 emetivi (= MTV; eng; proper noun)
 essebetê (= SBT; proper noun)
 Estúdio bê (proper noun)
 FAE (proper noun)
 fenemê (= FNM; proper noun)
 FIAT (proper noun)
 FUNASA (proper noun)
 FUNDEP (proper noun)
 FUNED (nome próprio)
 geemepecê (= GMPc; nome)
 GMar (nome próprio)
 Grupo á (= Grupo A; nome próprio)
 iemedebê (= IMDb; nome próprio)
 iemeele (= IML; nome próprio)
 ieneesseesse (= INSS; nome próprio)

ienepeesse (= INPS; nome próprio)
 iuessei (= USE; ing.; nome próprio)
 Letras-LIBRAS (nome próprio)
 LIBRAS (nome próprio)
 MAI (nome próprio)
 oquei (= OK, em “tá oquei”, “detergente oquei”, “papel toalha oquei”; adverb)
 oquei (= OK; adverb)
 pedeé cinco (= PDE5; noun + numeral)
 pedeeffe (= PDF; proper noun)
 pedevê (= PDV; proper noun)
 peefe (= PF; “prato feito”; noun)
 peerrepê (= PRP; noun)
 petê (= PT; proper noun)
 petebê (= PTB; proper noun)
 peueme (spelling of the word “pum”; noun)
 PUC (proper noun)
 PUC Minas (proper noun)
 RAI (proper noun)
 Rede tevê (proper noun)
 SAE (proper noun)
 SARS (proper noun)
 SENAC (proper noun)
 SENAI (proper noun)
 SERPRO (proper noun)
 SESC (proper noun)
 SUS (proper noun)
 teagaxis (name of movie; proper noun)
 teele (= TL; proper noun)
 teemepegenque (= TMPGEnc; proper noun)
 tejota (= TJ; proper noun)
 tevê (= TV; noun)
 teveal (= TVAL; proper noun)
 tevê Balcão (proper noun)
 tevê Colosso (proper noun)
 tevê Mais (proper noun)
 três dê (= 3D; numeral + noun)
 uefeemegê (= UFMG; proper noun)
 UFOP (proper noun)
 uteí (= UTI; proper noun)
 veelecê (= VLC; proper noun)
 Vila CEMIG (name of a neighbourhood; proper noun)
 xis ípsilon zê (in “princípios xis ípsilon zê”; adjective + adjective + adjective)
 xixisípsilon (name of movie; proper noun)

Forms of the verbs *estar*, *ir*, *vir*, *ter*, *poder* e *deixar*

po' (= pode; verb)
 tá (= está; verb)
 tamo (= estamos; verb)
 tamos (= estamos; verb)
 tão (= estão; verb)
 tar (= estar; verb)

taria (= estaria; verb)
tás (= estás; verb)
tava (= estava; verb)
tavam (= estavam; verb)
távamos (= estávamos; verb)
tavas (= estavas; verb)
teja (= esteja; verb)
tem (= tenho; verb)
teve (= esteve; verb)
tive (= estive; verb)
tiver (= estiver; verb)
tiverem (= estiverem; verb)
tivesse (= estivesse; verb)
tô (= estou; verb)
vamo (= vamos; verb)
vão (= vamos; verb)
vim (= vir; verb)
xá (= deixa; verb)

First person plural, not standard forms and reduced paradigm

a' aqui (= olha aqui; verb + adverb)
acabamo (= acabamos; verb)
achamo (= achamos; verb)
agradecemos (= agradecemos; verb)
a' lá (= olha lá; verb + adverb)
a' o (= olha o; verb + article)
a' os (= olha os; verb + article)
aprendemo (= aprendemos; verb)
arrumamo (= arrumamos; verb)
assinávamo (= assinávamos; verb)
atravessamo (= atravessamos; verb)
avi (= vi; verb)
avinha (= vinha; verb)
bebemo (= bebemos; verb)
beijamo (= beijamos; verb)
botemo (= botamos; verb)
chegamo (= chegamos; verb)
cheguemo (= chegamos; verb)
choramo (= choramos; verb)
colocamo (= colocamos; verb)
começamo (= começamos; verb)
comemo (= comemos; verb)
comemoramo (= comemoramos; verb)
compramo (= compramos; verb)
conhecemo (= conhecemos; verb)
consequimo (= conseguimos; verb)
contamo (= contamos; verb)
conversamo (= conversamos; verb)
corremo (= corremos; verb)
cortamo (= cortamos; verb)
deixamo (= deixamos; verb)

descansamo (= descansamos; verb)
 descemo (= descemos; verb)
 devemo (= devemos; verb)
 empurramo (= empurramos; verb)
 encontramo (= encontramos; verb)
 entramo (= entramos; verb)
 envem (= vem; verb)
 envinha (= vinha; verb)
 escolhemo (= escolhemos; verb)
 esquecemo (= esquecemos; verb)
 estamo (= estamos; verb)
 estudemo (= estudamos; verb)
 evem (= vem; verb)
 falamo (= falamos; verb)
 fazido (= feito, em “tinha que ter feito”; verb)
 ficamo (= ficamos; verb)
 fize (= fiz; verb)
 fizemo (= fizemos; verb)
 fomo (= fomos; verb)
 for (= formos; verb)
 fraga (3^a. p. s. v. “flagrar”, = flagra; verb)
 fragando (ger. v. “flagrar”, = flagrando; verb)
 frago (1^a. p. s. v. “flagrar”, = flagro; verb)
 fumo (= fomos; verb)
 ganhamo (= ganhamos; verb)
 levamo (= levamos; verb)
 levantamo (= levantamos; verb)
 levantemo (= levantamos; verb)
 mandamo (= mandamos; verb)
 manti (= mantive; verb)
 o' (= olha; verb or intejection)
 o' lá (= olha lá; verb + adverb)
 o' os (= olha os; verb + article)
 paramo (= paramos; verb)
 passamo (= passamos; verb)
 pedimo (= pedimos; verb)
 peguemo (= pegamos; verb)
 perdemo (= perdemos; verb)
 pinchando (= pichando; verb)
 pintemo (= pintemos; verb)
 podemo (= podemos; verb)
 precisamo (= precisamos; verb)
 pusemo (= pusemos; verb)
 resolvemo (= resolvemos; verb)
 saímo (= saímos; verb)
 seje (= seja; verb)
 sentamo (= sentamos; verb)
 sentemo (= sentamos; verb)
 separamo (= separamos; verb)
 somo (= somos; verb)
 sufro (= sofro; verb)
 temo (= temos; verb)

tiramo (= tiramos; verb)
tivemo (= tivemos; verb)
tomamo (= tomamos; verb)
trabalhamo (= trabalhamos; verb)
trago (= trazido, em “por nũ ter trago”; verb)
ver (= vir; fut. v. “ver”; verb)
vesse (= visse; verb)
viemo (= viemos; verb)
vimo (= vimos; verb)

Form *tó*

tó (= toma; verb)

Cliticization of subject pronoun

cê (= você; pronoun)
cês (= vocês; pronoun)
e' (= ele; pronoun)
ea (= ela; pronoun)
eas (= elas; pronoun)
es (= eles; pronoun)
ocê (= você; pronoun)
ocês (= vocês; pronoun)

Demonstratives pronouns

aque' (= aquele; demonstrative)
aquea (= aquela; demonstrative)
aqueas (= aquelas; demonstrative)
aques (= aqueles; demonstrative)

Prepositions

ca (= com a; preposition + article)
co (= com o; preposition + article)
cos (= com os; preposition + article)
cum (= com um; preposition + article)
cuma (= com uma; preposition + article)
d' a coisa (= de a coisa; preposition + article + nome)
d' a gente (= de a gente; preposition + article + nome)
d' aonde (= de aonde; preposition + adverb)
d' assim (= de assim; preposition + adverb)
d' o (= de o, in “apesar d' o ritual ser masculino”, “d' o Haroldo nũ tar buzinando”; preposition + article)
dum (= de um; preposition + article)
duma (= de uma; preposition + article)
dumas (= de umas; preposition + article)
duns (= de uns; preposition + article)
n' aonde (= onde; preposition + adverb)

n' deerrei (= na DRI; preposition + article + proper noun)
 ni (= em; preposition)
 n' onde (= onde; preposition + adverb)
 num (= em um; preposition + article)
 numa (= em uma; preposition + article)
 numas (= em umas; preposition + article)
 pa (= para, para a; preposition/preposition + article)
 pas (= para as; preposition + article)
 p' Belo Horizonte (= para Belo Horizonte; preposition + proper noun)
 p' Centro (= para o Centro; preposition + proper noun)
 p' chegar (= para chegar; preposition + verb)
 p' começar (= para começar; preposition + verb)
 p' falar (= para fazer; preposition + verb)
 p' fazer (= para fazer; preposition + verb)
 p' fritar (= para fritar; preposition + verb)
 p' pai (= para o pai; preposition + article + nome)
 p' pedir (= para pedir; preposition + verb)
 p' pessoal (= para o pessoal; preposition + noun)
 po (= para o; preposition + article)
 p' poder (= para poder; preposition + verb)
 pos (= para os; preposition + article)
 p' quando (= para quando, em "p' quando cê faz a massa", "p' quando a gente for usar"; preposition + adverb)
 pra (= para, para a; preposition/preposition + article)
 pr' aí (= para aí; preposition + adverb)
 pras (= para as; preposition + article)
 pro (= para o; preposition + article)
 pr' onde (= para onde; preposition + adverb)
 pros (= para os; preposition + article)
 prum (= para um; preposition + article)
 pruma (= para uma; preposition + article)
 pruns (= para uns; preposition + article)
 p' sair (= para sair; preposition + verb)
 p' São José do Norte (= para São José do Norte; preposition + proper noun)
 p' trazer (= para trazer; preposition + verb)
 pum (= para um; preposition + article)
 puma (= para uma; preposition + article)

Combinações de preposições e pronomes

c' aqueas (= com aquelas; preposition + demonstrative)
 c' aquela (= com aquela; preposition + demonstrative)
 c' cê (= com você; preposition + pronoun)
 c' e' (= com ele; preposition + pronoun)
 c' ele (= com ele; preposition + pronoun)
 c' essas (= com essas; preposition + demonstrative)
 c' esse (= com esse; preposition + demonstrative)
 c' ocê (= com você; preposition + pronoun)
 c' ocês (= com vocês; preposition + pronoun)
 daque' (= daquele; preposition + demonstrative)
 daquea (= daquela; preposition + demonstrative)
 daqueas (= daquelas; preposition + demonstrative)

daques (= daqueles; preposition + demonstrative)
 d' cê (= de você; preposition + pronoun)
 de' (= dele; preposition + pronoun)
 dea (= dela; preposition + pronoun)
 d' ela (= de ela; preposition + pronoun)
 d' ele (= de ele; preposition + pronoun)
 d' eles (= de eles; preposition + pronoun)
 des (= deles; preposition + pronoun)
 d' es (= de eles; preposition + pronoun)
 d' eu (= de eu; preposition + pronoun)
 d' ocê (= de você; preposition + pronoun)
 d' ocês (= de vocês; preposition + pronoun)
 naque' (= naquele; preposition + demonstrative)
 naquea (= naquela; preposition + demonstrative)
 naques (= naqueles; preposition + demonstrative)
 ne' (= nele; preposition + pronoun)
 n' ocê (= em você; preposition + pronoun)
 n' ocês (= em vocês; preposition + pronoun)
 p' aque' (= para aquele; preposition + demonstrative)
 p' aquele (= para aquele; preposition + demonstrative)
 p' cê (= para você; preposition + pronoun)
 p' cês (= para vocês; preposition + pronoun)
 p' e' (= para ele; preposition + pronoun)
 p' eu (= para eu; preposition + pronoun)
 p' ela(s) (= para ela(s); preposition + pronoun)
 p' ele (= para ele; preposition + pronoun)
 p' es (= para eles; preposition + pronoun)
 p' esse (= para esse; preposition + demonstrative)
 p' mim (= para mim; preposition + pronoun)
 p' ocê (= para você; preposition + pronoun)
 p' ocês (= para vocês; preposition + pronoun)
 pr' aquela (= para aquela; preposition + demonstrative)
 pr' aquilo (= para aquilo; preposition + demonstrative)
 pr' ea (= para ela; preposition + demonstrative)
 pr' ele (= para ele; preposition + demonstrative)
 pr' eu (= para eu; preposition + pronoun)
 pr' ocê (= para você; preposition + pronoun)
 pr' ocês (= para vocês; preposition + demonstrative)
 p' sio' (= para a senhora; preposition + pronoun)
 p' siora (= para a senhora; preposition + pronoun)
 p' sua (= para a sua; preposition + pronoun)

Plural marking

à vez (= às vezes; adverb)
 bícep (= bíceps; noun)
 Congonha (= Congonhas; proper noun)
 Esmeralda (= Esmeraldas; proper noun)
 Mina (= Minas; proper noun)
 Nossa Senhora das Graça (= Nossa Senhora das Graças; proper noun)
 Pato de Mina (= Patos de Minas; proper noun)
 Ribeirão das Neve (= Ribeirão das Neves; proper noun)

trícep (= tríceps; noun)

Exclamations to affirm or deny

ahn (interjection)
ahn ahn (interjection)
ham (interjection)
ham ham (interjection)
hum (interjection)
hum hum (interjection)
uhn (interjection)
uhn uhn (interjection)

Exclamations

eh (interjection)
ô (interjection)

Religious exclamations

Aff' (interjection)
No' (= Nossa; interjection)
Nossa Sio' (interjection)
Nossa Siora (interjection)
Nu' (= Nossa; interjection)
Nusga (= Nu', Nossa; interjection)
Vix' (interjection)
Vixe' (interjection)

Onomatopoeias

au (onomatopoeia)
bá bá bá bá bá bá (onomatopoeia)
blá blá blá blá blá (onomatopoeia)
brá (onomatopoeia)
lig lig lig lig lig lig (onomatopoeia)
pá (onomatopoeia)
pá pá (onomatopoeia)
pá pá pá (onomatopoeia)
piu (onomatopoeia)
piu piu (onomatopoeia)
pu (onomatopoeia)
puf (onomatopoeia)
puf pua (onomatopoeia)
su su su (onomatopoeia)
tã tã tã (onomatopoeia)
tá tá tum (onomatopoeia)
tec tec (onomatopoeia)
tê tô tó (onomatopoeia)
tic tic tic (onomatopoeia)
toc toc toc (onomatopoeia)
toque (onomatopoeia)

tuf (onomatopoeia)
tu tu (onomatopoeia)
ué (onomatopoeia for baby crying)
xim pá (onomatopoeia)

Negations

n' é (= não é; adverb + verb)
né (= não é; adverb + verb)
n' era (= não era; adverb + verb)
nũ (= não; adverb)

Apocopes

canarim (= canarinho(s); noun)
espinim (= espinho(s); noun)
padrim (= padrinho(s); noun)
passarim (= passarinho(s); noun)
porco-espinim (= porco-espinho; noun)
sozim (= sozinho, sozinha; adjective)

Diminutives

almoçozim (= almoçozinho; diminutive noun)
amarelim (= amarelinho; diminutive adjective)
azulzim (= azulzinho, in “prefiro o azulzim”; diminutive adjective)
bebezim (= bebezinho, in “peguei um bebezinho”; diminutive noun)
bichim (= bichinho; diminutive noun)
bocadim (= bocadinho; diminutive noun)
bonitim (= bonitinho; diminutive adjective)
cachorrim (= cachorrinho; diminutive noun)
cantim (= cantinho; diminutive noun)
capoeirim (= capoeirinhas; diminutive noun)
carrim (= carrinho; diminutive noun)
cedezinho (diminutivo de “CD”; diminutive noun)
certim (= certinho; diminutive adjective)
certins (= certinhos; diminutive adjective)
Chapeuzim Vermelho (= Chapeuzinho Vermelho; proper noun)
chazim (= chazinho; diminutive noun)
controladim (= controladinha; diminutive adjective)
desfiadim (= desfiadinho; diminutive adjective)
direitim (= direitinho, in “organizadinha / tudo direitim”; diminutive adjective)
direitim (= direitinho, in “cuida direitim”, “arruma direitim”; adverb)
esquisitim (= esquisitinho; diminutive adjective)
fechadim (= fechadinho; diminutive adjective)
filhotim (= filhotinho; diminutive noun)
formulariozim (= formulariozinho; diminutive noun)
fundim (= fundinho; diminutive noun)
Geraldinim (= Geraldinho; proper noun)
golezim (= golezinho; diminutive noun)
igualzim (= igualzinho, in “igualzim um Big Brogher”, “igualzim de casa de pobre”; conjunction (diminutiva?))

instantim (= instantinho; diminutive noun)
jeitim (= jeitinho; diminutive noun)
Joãozim (= Joãozinho; proper noun)
joguim (= joguinho; diminutive noun)
ladim (= ladinho; diminutive noun)
maciim (= maciinho; diminutive adjective)
mansim (= mansinho; diminutive adjective)
Marquim (= Marquinho; proper noun)
meninim (= menininho; diminutive noun)
morenim (= moreninho, in “mais morenim”; diminutive adjective)
murim (= murinho; diminutive noun)
Paulim (= Paulinho; proper noun)
pequeninim (= pequenininha; diminutive adjective)
pertim (= pertinho; adverb)
negocim (= negocinhos; diminutive noun)
partidim (= partidinho; diminutive adjective)
porquim (= porquinho; diminutive noun)
portim (= portinha; diminutive noun)
potim (= potinho; diminutive noun)
pouquim (= pouquinho, in “um pouquim”; diminutive noun)
pozim (= pozinho; diminutive noun)
pretim (= pretinho; diminutive adjective)
prontim (= prontinho; diminutive adjective)
quadradim (= quadradinha; diminutive adjective)
quadradim (= quadradinho; diminutive adjective)
queimadim (= queimadinho; diminutive adjective)
rapidim (= rapidinho; adverb)
retheadim (= retheadinho; diminutive adjective)
rolim (= rolinho; diminutive noun)
tamanim (= tamaninho; diminutive noun)
tampadim (= tampadinho; diminutive adjective)
terrenim (= terreninho; diminutive noun)
tiquim (= tiquinho; diminutive noun)
todim (= todinho, in “ele molha todim”; adverb)
toquim (= toquinho; diminutive noun)
trancadim (= trancadinhos; diminutive adjective)
trenzim (= trenzinho; diminutive noun)
tudim (= tudinho; pronoun (diminutive))

Senhor e senhora

seu (= senhor; pronoun)
sio' (= senhora; pronoun)
sior (= senhor; pronoun)
siora (= senhora; pronoun)
sô (= senhor; pronoun)

Intendifier mó

mó (= maior, in “mó amor/intimidade/confusão/palha/gritão”; adjective)
mó (= muito, in “mó bonitinha/gostoso/feliz/bom”; adverb)

Rotacism

armoçar (= almoçar; verb)
artinho (= alinho, in “do artinho assim eu vi e’ ”; diminutive noun)
arto (= alto, in “no lugar mais arto”; adjective)
arto (= alto, in “aqui no arto” e “viu ela do arto”; noun)
comprica (= complica; verb)
compricar (= complicar; verb)
cravícula (= clavícula; noun)
escardada (= escaldada; adjective)
prano (= plano; noun)
pranta (= planta; noun)
pray (= play, in “dá o pray”, “liga o pray”; ing.; noun or verb)
prissado (= plissado; noun)
problemas (= problemas; noun)
sortando (= soltando; verb)
sortar (= soltar; verb)
sortei (= soltei; verb)
sorto (1ª. p. s. v. “soltar”, = solto; verb)
sortou (= soltou; verb)
vorta (3ª. p. s. v. “voltar”, = volta; verb)
vortar (= voltar; verb)
vortava (= voltava; verb)
vorto (= volto; verb)

Readings

nanananã (substitutes not interesting reading parts)

Others

etcetera (= et cetera; lat.; conjunction)
&he (time taking)
xxx (not understandable word)
yyy (censored word)
yyyy (not understandable part (more than one word))

APHERETIC FORMS

babacar (= embabacar; verb)
baixa (= abaixa; verb)
baixar (= abaixar; verb)
baixei (= abaixei; verb)
baulado (= abaulado; adjective)
bora (= embora; adverb)
borrecido (= aborrecido; adjective)
brigada (= obrigada; adjective)
brigado (= obrigado; adjective)
caba (= acaba; verb)
cabar (= acabar; verb)
cabava (= acabava; verb)

cabei (= acabei; verb)
 cabou (= acabou; verb)
 celera (= acelera; verb)
 celerando (= acelerando; verb)
 certar (= acertar; verb)
 chei (= achei; verb)
 cho (= acho; verb)
 contece (= acontece; verb)
 contecer (= acontecer; verb)
 conteceu (= aconteceu; verb)
 cordava (= acordava; verb)
 creditei (= acreditei; verb)
 dianta (= adianta; verb)
 doro (= adoro; verb)
 dotada (= adotada; adjective)
 fessora (= professora; noun)
 final (= afinal; adverb)
 fundar (= afundar; verb)
 garrado (= agarrados; adjective)
 garrou (= agarrou; verb)
 gateelevê (= agateelevê (HTLV); proper noun)
 gora (= agora; adverb)
 gual (= igual, in “gual luís-cacheiro”; conjunction)
 gualzim (= igualzinho, in “gualzim lá em casa”; conjunction (diminutive))
 güenta (= agüenta; verb)
 güentando (= agüentando; verb)
 güentar (= agüentar; verb)
 güento (= agüento; verb)
 güentou (= agüentou; verb)
 inda (= ainda; adverb)
 judar (= ajudar; verb)
 lambique (= alambique; noun)
 laranjado (= alaranjado, in “as cores / muito / laranjado”; adjective)
 lisou (= alisou; verb)
 magina (= imagina; verb)
 mamentar (= amamentar; verb)
 manhã (= amanhã; adverb)
 marelo (= amarelo, in “cacho marelo”; adjective)
 marrava (= amarrava; verb)
 migão (= amigão; augmentative noun)
 mor (= amor; noun)
 ném (= neném; noun)
 panhava (= apanhava; verb)
 parece (= aparece; verb)
 pareceu (= apareceu; verb)
 partamento (= apartamento; noun)
 pelido (= apelido; noun)
 pera (= espera; verb)
 perta (= aperta; verb)
 pertar (= apertar; verb)
 pertei (= apertei; verb)
 pesar (= apesar; adverb)

pinhada (= apinhada; adjective)
 pois (= depois; adverb)
 posa (= raposa; noun)
 proveita (= aproveita; verb)
 proveitando (= aproveitando; verb)
 proveitei (= aproveitei; verb)
 purra (= empurra; verb)
 qui (= daqui; preposition + adverb)
 rancaram (= arrancaram; verb)
 rancava (= arrancava; verb)
 rancou (= arrancou; verb)
 ranjar (= arranjar; verb)
 ranjasse (= arranjasse; verb)
 ranjou (= arranjou; verb)
 rebentando (= arrebetando; verb)
 rebentar (= arrebetar; verb)
 regaço (= arregaços; noun)
 rorosa (= horrorosa; adjective)
 roz (= arroz; noun)
 rumaram (= arrumaram; verb)
 sobiando (= assobiando; verb)
 tá (= está; verb)
 tadim (= tadinho; interjection (diminutive))
 tadinha (= coitadinha; interjection (diminutive))
 tadinho(s) (= coitadinho(s); interjection (diminutive))
 tamo (= estamos; verb)
 tamos (= estamos; verb)
 tão (= então; adverb)
 tão (= estão; verb)
 tar (= estar; verb)
 taria (= estaria; verb)
 tás (= estás; verb)
 tava (= estava; verb)
 tavam (= estavam; verb)
 távamos (= estávamos; verb)
 tavas (= estavas; verb)
 té (= até, in “té onça”, “posso té ir”, “té / né / a gente trabalha”, “eu té avi as defesa”; adverb)
 té (= até, in “té hoje nũ deu em nada”; preposition)
 teirinho (= inteirinho; diminutive adjective)
 teja (= esteja; verb)
 tendeu (= entendeu; verb)
 tendi (= entendi; verb)
 testino (= intestino; noun)
 teve (= esteve; verb)
 tive (= estive; verb)
 tiver (= estiver; verb)
 tiverem (= estiverem; verb)
 tivesse (= estivesse; verb)
 tô (= estou; verb)
 tradinha (= entradinha; diminutive noun)
 trapalha (= atrapalha; verb)
 trapalhado (= atrapalhado; adjective)

trapalhou (= atrapalhou; verb)
travessa (= atravessa; verb)
travessadinho (= atravessadinho; diminutive adjective)
trevidão (= atrevidão; augmentative adjective)
vó (= avó; noun)
vô (= avô; noun)
xá (= deixa; verb)

OTHER FORMS

abstratismo (noun)
acampantes (aqueles que acampam; noun)
aceto balsâmico (noun)
ácido bode (= ácido bórico; noun)
ácido bóide (= ácido bórico; noun)
add (ing.; verb)
aê (interjection)
aftermarket (ing.; noun)
air bags (ing.; noun)
airutu (probabçy for “urutu”; noun)
ajuadaria (= ajudaria; verb)
al dente (it.; adjective)
alentejama (= alentejana; adjective)
aletejana (= alentejana; adjective)
alunos-problema (noun)
alunos-problemas (noun)
amava (= amável; adjective)
anche (it.; conjunction)
antonte (= anteontem; adverb)
apeludo (= apelido; noun)
a priori (lat.; adverb)
arbetura (= abertura; noun)
arco (= álcool; noun)
arrima (= arrimo; noun)
arruim (= ruim; adjective)
artisticando (verb)
art nouveau (fr.; noun)
astubóide (= ácido bórico; noun)
atelier (fr.; noun)
au (onomatopoeia)
autora (in “a gente autora um devedê”; verb)
a ver (esp.; verb)
avi (= vi; verb)
avinha (= vinha; verb)
bá bá bá (for concluding a list, like “et cetera”)
bá bá bá bá bá bá (onomatopoeia)
babies (eng.; noun)
bandeira (= tamanduá-bandeira; noun masculine)
Barbie Guel (= Barbie Girl; proper noun)
barrançudo (adjective)
barreta (= barrete; noun)

basicon (aumentativo de “básica”; augmentative adjective)
 batcaverna (noun)
 batidim (= batidinhas, in “c’ aqueas batidim pesada de’ ”; noun diminutive)
 bebão (= bebadozão; augmentative adjective)
 because (eng.; conjunction)
 bicepsinho (diminutivo de “bíceps”; noun diminutive)
 big (eng.; adjective)
 Big Brogher (= Big Brother; proper noun)
 bisote (in “acabamento bisote”; adjective)
 bisotear (verb)
 blá blá blá blá blá (onomatopoeia)
 bleh (exprime nojo; interjection)
 blue (in “cê quer o blue”; eng.; adjective)
 blue jeans (eng.; noun)
 blueszão (aumentativo de “blues”; eng.; augmentative noun)
 boaça (aumentativo de “boa”, in “tava de boaça”; augmentative adjective)
 body art (eng.; noun)
 boiado (in “ele tava boiado” (= boiando); adjective)
 bombou (v. “bombar”; verb)
 Borbagato (= Borbagatur; proper noun)
 Borges das Costa (= Borges da Costa; proper noun)
 brá (onomatopoeia)
 breja (= cerveja; noun)
 brise (fr.; noun)
 bro (= brother; eng.; noun)
 bubu (= chupeta, bico; noun)
 buffet (fr.; noun)
 busão (= ônibus; augmentative noun)
 bye-bye (eng.; interjection)
 caça-talents (noun)
 cachaceira-mor (noun)
 calcanha (= calcanhar; noun)
 canarina (fem. de “canarinho”; noun)
 cappelletti (it.; noun)
 capu (= capô; noun)
 caraca (interjection)
 caracangaia (noun)
 cara-de-paumente (adverb derived from “cara-de-pau”; adverb)
 careful (eng.; adjective)
 carquer (= qualquer, in “carquer lugar tá bom”; adjective)
 carreado (= encarreirado; verb, participle)
 cascalhada (de cascalho, com cascalho, in “a rodovia era cascalhada”; adjective)
 casona (augmentative of “casa”; augmentative noun)
 cato (= quatro; numeral)
 cerurgia (= cirurgia; noun)
 chat (eng.; noun)
 cheddar (eng.; noun)
 cheddar (= cheddar; eng.; noun)
 ciminha (diminutive of “cima”, in “em ciminha”; diminutive noun)
 coisica (= coisinha; diminutive noun)
 come stai (it.; adverb + verb)
 come tu ti chiami (it.; adverb + pronoun + pronoun + verb)

come with me (eng.; verb + preposition + pronoun)
 comieira (port. eur.; noun)
 comment (ing; noun)
 completed (ing; adjective)
 condômino (= condomínio; noun)
 confirações (= configurações; noun)
 constratarem (= contrataram; verb)
 corpus (lat.; noun)
 country (in “dança country”; eng.; adjective)
 cover (in “a gente faz um cover”; eng.; noun)
 cozindo (in “cozindo as trouxas e os baús de frande”; verb)
 crack (eng.; noun)
 crebrei (= quebrei; verb)
 cronofone (= cronofone; noun)
 crossover (eng.; noun)
 cult (eng.; adjective)
 d’água (= de água; preposition + noun)
 datashow (noun)
 dear (= dar; verb)
 dedeira (= mamadeira; noun)
 default (in “a pasta default dele”; ing; adjective)
 default (in “tem um default dele”, “vou deixar default mesmo”; eng.; noun)
 delicious (eng.; adjective)
 deromou (= demorou; verb)
 designers (eng.; noun)
 desktop (eng.; noun)
 desmexeu (verb)
 desnegocar (verb)
 depois (= depois; adverb)
 diavolo (it.; noun)
 diet (eng.; adjective)
 dizar (verb)
 dócia (= dócil; adjective)
 Doctor nine ou two one ou (proper noun)
 doidado (augmentative of “doido”; augmentative adjective)
 don’t use (eng.; verb + adverb + verb)
 drag (eng.; noun)
 drag queens (eng.; noun)
 edit (eng.; verb)
 eight (eng.; numeral)
 e-mail (eng.; noun)
 embromation (= embromação; noun)
 emo (noun)
 encarfunado (adjective)
 enfermeira-chefe (noun)
 entregou (= entregou; verb)
 enter (eng.; verb)
 entrilhado (adjective)
 envem (= vem; verb)
 envinha (= vinha; verb)
 escoveu (= escolheu; verb)
 escrotando (derived from “escroto”; verb)

espinhãozinho (in “espinhãozinho de ostra”; diminutive noun)
 esporrodando (verb)
 espumé (game word; noun?)
 est (lat.; verb)
 estrupição (= estrupício; noun)
 estrupo (= estrupo; noun)
 evem (= vem; verb)
 facinho (diminutive of “fácil”; diminutive adjective)
 falazada (noun)
 farmaco (reduction of “farmacologia” or of word meaning a different discipline of Pharmacology
 faculty; noun)
 fazeção (noun)
 fazido (= feito, in “tinha que ter fazido”; verb)
 feedback (eng.; noun)
 feeling (eng.; noun)
 file (eng.; noun)
 fize (= fiz; verb)
 flash (eng.; noun)
 flashback(s) (eng.; noun)
 fluoxetina (noun)
 fofocalhada (derivation like “brigalhada”; pronounced “fofocaiada” and “brigaiada”; noun)
 foicinha (dim. de “foice”; diminutive noun)
 fondue (fr.; noun)
 format (eng.; verb)
 fosfro (= fósforo; noun)
 fotinha (diminutivo de “foto”; diminutive noun)
 fraga (3^a. p. s. v. “flagrar”, = flagra; verb)
 fragando (ger. v. “flagrar”, = flagrando; verb)
 frago (1^a. p. s. v. “flagrar”, = flagro; verb)
 frande (in “cozindo as trouxas e os baús de frande”; noun?)
 freelance (eng.; adverb)
 freezer (eng.; noun)
 frentona (aumentativo de “frente”; augmentative noun)
 friends (eng.; noun)
 frosco (= fósforo; noun)
 funk (eng.; noun)
 funkão (augmentative noun)
 ganso-açu (noun)
 gay (in “comunidade gay”, “muito gay”, “fica gay”, “que gay”, “muito gay”, “era gay”; eng.;
 adjective)
 gays (in “dois gays”; eng.; noun)
 glandão (= grandão; augmentative adjective)
 gloss (eng.; noun)
 go for it (eng.; verb + preposition + pronoun)
 go-go boy (eng.; noun)
 grada (in “ostra grada, “pra ficar grada”; adjective)
 guesingnação (= designação; noun)
 ha (interjection)
 ha ha (interjection)
 han (termo de jogo; noun)
 handsome (eng.; adjective)
 happening (eng.; noun)

haqueou (der. de “hack” (eng.); verb)
 hard-core (in “tocar um hard-core nacional”, “curtir um hard-core”; eng.; noun)
 hay que endurecerse sin perder la ternura jamás (esp.; verb + conjunction + verb + pronoun + preposition + verb + artigo + noun + adverb)
 hein (interjection)
 Heinekente (Word made by the words “Heineken” and “quente”; proper noun)
 hello (eng.; interjection)
 hobby (eng.; noun)
 homem-cobra (noun)
 homesnagedos (= homenageados; adjective)
 homework (eng.; noun)
 homogeniza (3^a. p. s. v. “homogeneizar”, = homogeneiza; verb)
 hot (eng.; adjective)
 hu hu (interjection)
 hype (eng.; noun)
 I don’t give my heart to one ’cause I don’t wanna waste my time (eng.; pronoun + verb + adverb + verb + pronoun + noun + preposition + pronoun + conjunction + pronoun + verb + adverb + verb + preposition + verb + pronoun + noun)
 iens (game word; noun)
 imagination (eng.; noun)
 infelizio (= infeliz; adjective)
 interfamiliares (adjective)
 internet (eng.; noun)
 io mi chiamo (it.; pronoun + pronoun + verb)
 iorgute (= iogurte; noun)
 irai (noun)
 jataí (noun)
 jeans (in “o poder duma calça jeans”; eng.; adjective)
 joão-urutu (noun)
 juninha (= neófito; noun)
 kami (= papel (part of the original word for “origami”); jap.; noun)
 karowara (ind.; noun)
 ketchup (eng.; noun)
 kit (eng.; noun)
 kitsch (ger.; noun)
 kong (game word; noun?)
 la hermana (sp.; artigo + noun)
 laptop (eng.; noun)
 latex (= látex; noun)
 lato sensu (lat.; adverb)
 Lei (it.; pronoun)
 lettering(s) (eng.; noun)
 level (eng.; noun)
 light (eng.; adjective)
 lig lig lig lig lig lig lig (onomatopéia)
 line out (in “isso aqui é o line out”; eng; noun)
 link (eng.; noun)
 logos (reduction of “logomarcas”; noun)
 londeira (= ladeira; noun)
 looping (in “sete minutos com o looping”; eng.; noun)
 love (in “um love”; eng.; noun)
 madeirinho (diminutive noun)

madonnice (noun)
 majongue (noun)
 majors (eng.; noun)
 maleira (in “ela é mó maleira”; adjective)
 mamazinho (diminutive noun)
 mandaçaia (noun)
 mano de obra (sp.; noun (ou: noun + preposition + noun?))
 mano yo (sp.; noun + pronoun)
 manti (= mantive; verb)
 mantinha (diminutive of “manta”; diminutive noun)
 marchand (fr.; noun)
 margarita (= margherita; it.; noun)
 margherita (it.; noun)
 markup (eng.; noun)
 maroca (noun)
 massiva (in “tecnologia massiva”; adjective)
 max (tyoe of pizza; noun)
 meandar (= mandar; verb)
 media output (eng.; noun (ou adjective + noun?))
 menu (fr.; noun)
 mercânica (= mecânica, in “perna mercânica”; adjective)
 merchand (probably reduction of “merchandising”; eng.; noun)
 mic (in “ “mic” de microfone”; explaining that “mic” means “microfone”; noun)
 micrim (diminutive of “microondas”; diminutive noun)
 miniminiza (= minimiza; verb)
 mise-en-scène (fr.; noun)
 modus operandi (lat.; noun)
 Moelhus (= Moebius; proper noun)
 moletim (tipo de moletom; noun)
 mono (reduction of “monofonia” or of “monofônico”, in “programação pra mono”, “tá escrito “mono” ”, “tá gravando em mono”; noun ou adjective?)
 monologue (eng.; noun)
 mota (= moto; noun)
 motinha (diminutivo de “moto”; diminutive noun)
 mouse (eng.; noun)
 much better (eng.; adverb + adjective)
 muquifo (noun)
 music (eng.; noun)
 my computer (eng.; pronoun + noun)
 nananã (like “et cetera”)
 nanananã (substitutes reading parts that the reader does not judge interesting to be read)
 nana nana ã ã ã (substitutes parto f songs that the Singer does not know how to sing)
 não-comprimido (adjective)
 navy (indicating purse decoration pattern; eng.; adjective)
 negão (= negrão; augmentative noun)
 nego (= negro; noun)
 negociando (verb)
 ném (= neném; noun)
 nerd (eng.; noun)
 network (eng.; noun)
 non aedificandi (lat.; verb)
 nortelos (game word; noun?)

not (eng.; adverb)
 notebook (eng.; noun)
 Nusga (= Nu', Nossa; interjection)
 ó (slang: "a festa tava o ó"; "é o ó do borogodó"; noun)
 ocurujal (noun)
 off (in "em off"; eng.; adverb?)
 omnidirecional (adjective)
 onça-café (noun)
 ônibu (= ônibus; noun)
 ôniu (= ônibus; noun)
 online (in "tá online"; eng.; adjective)
 oops (eng.; interjection)
 open (eng.; verb)
 óptemo (= ótimo; adjective)
 ori (= dobra (parto f the original Word for "origami"); jap.; noun ou verb?)
 origami (jap.; noun)
 outono-inverno (in "coleção outono-inverno"; adjective)
 output video with high (eng.; noun (ou adjective + noun?) + preposition + adjective)
 ora-pro-nóbi (= ora-pro-nóbis; noun)
 pá (onomatopoeia)
 paella (sp.; noun)
 paella marinera (sp.; noun + adjective)
 paiaço (made by the two words "pai" and "palhaço"; noun)
 paiê (allocutive for "pai"; noun)
 pá pá (onomatopoeia)
 pá pá pá (onomatopoeia)
 paparito(s) (noun)
 papel-toalha (noun)
 pause (in "cê dá um pause aqui"; eng.; noun)
 pay-per-view (eng.; noun)
 pedacico (= pedacinho; diminutive noun)
 peguete (noun)
 pen drive (eng.; noun)
 perhaps love is like the ocean (part of a song; eng.; adverb + noun + verb + conjunction + artigo + noun)
 permesso di soggiorno (it.; noun + preposition + noun)
 perolato (in "sombra branca cintilante // perolato // perolada //"; adjective ou noun?)
 personal killer (eng.; adjective + noun)
 personal trainer (eng.; adjective + noun)
 pinchando (= pichando; verb)
 piripaque (noun)
 piu (onomatopoeia)
 piu piu (onomatopoeia)
 pixota (in "ele é pixota demais"; adjective)
 pizza (it.; noun)
 pizzaiola (it.; noun)
 play (in "dá o play aí"; eng.; noun)
 playboy (in "corte playboy"; eng.; adjective)
 pô (interjection)
 problema (= problema; noun)
 pobrema (= problema; noun)
 pong de dora (game word; noun + preposition + noun)

pong de dragão (game word; noun + preposition + noun)
 ponsto (= pontos; noun)
 ponto-paris (noun)
 pop (eng.; adjective)
 pop art (eng.; noun)
 português-inglês (major in faculty of Letras; noun)
 pós (= pós-graduação; noun)
 posquim (= porquinhos; diminutive noun)
 posquinho (= porquinhos; diminutive noun)
 profunda (= profundas; noun)
 prendeção (noun)
 preset (in “configuramos o preset três”; eng.; noun)
 prota (= prótese; noun)
 psiquiátricas (= psiquiatras; noun)
 pu (onomatopoeia)
 puf (onomatopoeia)
 puf pua (onomatopoeia)
 pure (it.; conjunction)
 putz (interjection)
 putz grila (interjection)
 quatros (= quatro; numeral)
 quorum (lat.; noun)
 rapaizi (= rapaz; noun)
 rapidão (augmentative adjective)
 rapim (noun)
 arroz (= arroz; noun)
 ready (eng.; adjective)
 reba (= riba; noun)
 rec (in “aperta o rec”; eng.; noun)
 receiver(s) (eng.; noun)
 refri (= refrigerante; noun)
 remote (ing; noun)
 retardadice (noun)
 Retrato de Ambrosie Vollad (= Retrato de Ambroise Vollard; proper noun)
 réveillon (fr.; noun)
 review (eng.; noun)
 ricona (aumentativo de “rica”; augmentative adjective)
 rock-'n'-roll (eng.; noun)
 rosto a rosto (expression of Alberto Roberto, in Chico Anysio's show; noun + preposition + noun)
 sagiu (= surgiu; verb)
 saria (= seria; verb)
 save (eng.; verb)
 script (eng.; noun)
 senuca (= sinuca; noun)
 sertralina (noun)
 set (eng.; noun)
 settings (eng.; noun)
 sexy (eng.; adjective)
 shift (in “tem que usar esse shift”; eng.; noun)
 shit (eng.; interjection)
 shopping (eng.; noun)
 short (in “short story/monologue”; eng.; adjective)

short (peça de roupa; eng.; noun)
 shortinho (diminutive noun)
 show(s) (eng.; noun)
 showzinho (diminutive noun)
 shoyu (jap.; noun)
 sibutramina (noun)
 siclano (= sicrano; noun)
 sildenafil (noun)
 sinfronismo (noun)
 site (eng.; noun)
 spa (eng.; noun)
 squash (eng.; noun)
 squashzinho (diminutive noun)
 start (eng.; verb)
 status (lat.; noun)
 sto bene (it.; verb + adverb)
 sto male (it.; verb + adverb)
 stop (in “eu dei stop”; eng.; noun ou verb?)
 story (eng.; noun)
 stricto sensu (lat.; adverb)
 sufro (= soffro; verb)
 super (adverb)
 superbaratinho (adjective diminutivo)
 superbarato (adjective)
 superbem (adverb)
 superbem-aceito (adjective)
 superbem-arrumada (adjective)
 superbofe (noun)
 superbonita (adjective)
 superbonitinha (adjective diminutivo)
 superdiscreta (adjective)
 superespecífico (adjective)
 superestourado (adjective)
 supergente boa (adjective)
 superinformal (adjective)
 superlegal (adjective)
 supermacho (noun)
 supernerd (noun)
 supernova (adjective)
 superpoderes (noun)
 superpreconceito (noun)
 superpreocupado (adjective)
 supersensível (adjective)
 su su su (onomatopoeia)
 tã dã dã (onomatopoeia)
 tadalafila (noun)
 tã tã tã (onomatopoeia)
 tá tá tum (onomatopoeia)
 tauba (= tábu; noun)
 tec tec (onomatopoeia)
 telemarketing (eng.; noun)
 tempro (= tempo; noun)

tês (= três; numeral)
 tesoro (it.; noun)
 tê tô tó (onomatopoeia)
 tic tic tic (onomatopoeia)
 tilte (= falha; noun)
 tinhas (wrong pronunciation for “tinha”, 3^a. p. s.; verb)
 tiquetaque (= prendedor de cabelo; noun)
 to (eng.; preposition)
 TOC (= Transtorno Obsessivo-Compulsivo; proper noun)
 toca (= troca; verb)
 tocar (= trocar; verb)
 toc toc toc (onomatopoeia)
 toím (= ânus; noun)
 top hit (eng.; adjective + noun)
 toque (onomatopoeia)
 town house (= town houses; eng.; noun)
 track erase (eng; noun + verb)
 trade (eng.; noun)
 tragalada (noun)
 trago (= trazido, in “por nũ ter trago”; verb)
 trailer (eng.; noun)
 transe (= trânsito; noun)
 transmitter(s) (eng.; noun)
 trash (eng.; noun)
 tri (in “também / umas coisa que eu nũ sei pra quê // espeto // tri //” (it is not onomatopoeia or reduction for “tricampeão/tricampeonato”, and does not seem interrupted word; noun?)
 trocentas (in “trocentas coisas”; numeral)
 tsuru (jap.; noun)
 tuf (onomatopoeia)
 tu tu (onomatopoeia)
 tutura (in “mexe com a minha tutura”; noun)
 two (eng.; numeral)
 ude (provavelmente, sigla, in “ude é o quê (question) Uberlândia (answer)”; noun)
 ué (onomatopoeia for baby crying)
 ué (interjection)
 uê (interjection)
 uf (= ufa; interjection)
 underground (in “som/garçom underground”; eng.; adjective)
 update (eng.; noun)
 upload (eng.; noun)
 urbanóide (adjective ou noun?)
 vale-tiquetezinhos (diminutivo de “vale-tíquetes”; diminutive noun)
 vardenafila (noun)
 ver (= vir; fut. v. “ver”; verb)
 vesse (= visse; verb)
 video output (eng.; noun (ou adjective + noun?))
 vintão (augmentative of “vinte”, in “o meu vintão”; augmentative noun)
 VIP (eng.; adjective)
 vivendos (= vivendo; verb)
 voi (it.; pronoun)
 vossuncê (= vosmecê; pronoun)
 vous (fr.; pronoun)

wafer (in “biscoito Mabel wafer”; eng.; adjective)
 watch and learn (eng.; verb + conjunction + verb)
 web (eng.; noun)
 well (in “well / vamos falar sobre musicologia”, “bom // well //”; eng.; interjection)
 white balance (eng.; adjective + noun)
 xim pá (onomatopoeia)
 xisinho (diminutivo de “xis” (x); diminutive noun)
 yes (eng.; interjection)
 zap (truco’s card; noun)
 zê-mandado (noun)
 zenes (game word; noun)
 zorelha (noun)
 zum (= zoom (eng.); noun)

Spelled words and expressions:

agabeesse (= HBs; part of “agabeesseagê” segmented; proper noun)
 agê (= Ag; part of “agabeesseagê” segmented; proper noun)

alen (part of “alentejana” segmented; adjective)
 tejana (part of “alentejana” segmented; adjective)

alete (part of “aletejana” (= alentejana) segmented; adjective)
 jana (part of “aletejana” (= alentejana) segmented; adjective)

ca (part of “capeta” silabada; noun)
 pe (part of “capeta” silabada; noun)
 ta (part of “capeta” silabada; noun)

colo (part of “coloca” segmented; verb)
 ca (part of “coloca” segmented; verb)

de (part of “depois” silabada; adverb)
 pois (part of “depois” silabada; adverb)

efeele (= FLV; part of “efeelevê” segmented; proper noun)
 vê (= FLV; part of “efeelevê” segmented; proper noun)

eme (= MPEG; part of “emepegue” segmented; proper noun)
 pegue (= MPEG; part of “emepegue” segmented; proper noun)

gela (part of “geladeira” segmented; noun)
 deira (part of “geladeira” segmented; noun)

micro (part of “microgotinhas” segmented; noun)
 gotinhas (part of “microgotinhas” segmented; noun)

Ita (part of “Itatiaia” segmented; proper noun)
 tiaia (part of “Itatiaia” segmented; proper noun)

nun (part of “nunca” spelled; adverb)
 ca (part of “nunca” spelled; adverb)

omidi (part of “omidieredicional” (= omnidirecional) segmented; adjective)
eredicional (part of “omidieredicional” (= omnidirecional) segmented; adjective)

pen (part of “pensando” segmented; verb)
sando (part of “pensando” segmented; verb)

Pla (part of “Placelar” segmented; proper noun)
celar (part of “Placelar” segmented; proper noun)

po (part of “população” segmented; noun)
pulação (part of “população” segmented; noun)

psico (part of “psicossomática” segmented; adjective)
somática (part of “psicossomática” segmented; adjective)

so (part of “somente” segmented; adverb)
mente (part of “somente” segmented; adverb)

super (part of “supereuropeísta” segmented; adjective)
europeísta (part of “supereuropeísta” segmented; adjective)

super (part of “superinteligente” segmented; adjective)
inteligente (part of “superinteligente” segmented; adjective)

ti (part of “tijolo” segmented; noun)
jolo (part of “tijolo” segmented; noun)

to (parte da expressão “toda vez” spelled; pronoun)
da (parte da expressão “toda vez” spelled; pronoun)
vez (parte da expressão “toda vez” spelled; noun)

Wiki (part of “Wikipedia” segmented; proper noun)
pedia (part of “Wikipedia” segmented; proper noun)

ze (part of “zerinho” silabada; numeral (diminutive))
ri (part of “zerinho” silabada; numeral (diminutive))
nho (part of “zerinho” silabada; numeral (diminutive))

Appendix 4 C-ORAL-BRASIL Prosodic Tagging Evaluation Report

1. Pre-validation (before the corpus prosodic tagging)

First test:

Segmentation of an 822 word dialogue and an 855 word monologue.

Results:

Group 1 Kappa – test 1

Agreement type	total	dl	mn
General agreement	0.79	0.83	0.75
Terminal breaks	0.83	0.90	0.74
Non-terminal breaks	0.61	0.62	0.61
No break	0.86	0.87	0.84

Group 1 Percentage– test 1

Agreement type	dl	mn
Total agreement	83%	91%
Terminal breaks	16%	8%
Non-terminal breaks	6%	8%
No break	67%	67%
Partial agreement	10%	15%
Terminal break vs. non-terminal	3.6%	6.9%
Non-terminal break vs. no break	6.7%	8.0%
Total disagreement	1.0%	1.4%
Terminal break vs. no break	0.6%	1.1%
Terminal break vs. non-terminal vs. no break	0.4%	0.4%

Second test:

Segmentation of a 719 word dialogue and a 784 word monologue.

Results:

Group 1 Kappa – test 2

Agreement type	total	dl	mn
General agreement	0.82	0.80	0.84
Terminal breaks	0.85	0.85	0.85
Non-terminal breaks	0.71	0.62	0.77
No break	0.88	0.87	0.89

Group 1 Percentage – test 2

Agreement type	dl	mn
----------------	----	----

Total agreement	84%	91%
Terminal breaks	15%	9%
Non-terminal breaks	7%	15%
No break	63%	65%
Partial agreement	13%	11%
Terminal break vs. non-terminal	5.7%	4.1%
Non-terminal break vs. no break	7.6%	6.9%
Total disagreement	0.8%	0.4%
Terminal break vs. no break	0.6%	0.1%
Terminal break vs. non-terminal vs. no break	0.3%	0.3%

Third test:

Segmentation of a 678 word dialogue and a 415 word monologue.

Results:

Group 1 Kappa – test 3

Agreement type	total	dl	mn
General agreement	0.77	0.78	0.76
Terminal breaks	0.82	0.87	0.71
Non-terminal breaks	0.62	0.58	0.66
No break	0.85	0.84	0.86

Group 1 Percentage – test 3

Agreement type	dl	mn
General agreement	86%	93%
Terminal breaks	13%	6%
Non-terminal breaks	7%	13%
No break	65%	65%
Partial agreement	15%	16%
Terminal break vs. non-terminal	5.2%	7.5%
Non-terminal break vs. no break	9.9%	8.7%
Total disagreement	0.3%	0.5%
Terminal break vs. no break	0.1%	0.2%
Terminal break vs. non-terminal vs. no break	0.1%	0.2%

Realistic Kappa Group 1 - pre-validation

Text	overall	terminal	non-terminal
Dialogue 1	0.59	0.81	0.50
Dialogue 2	0.56	0.76	0.47
Dialogue 3	0.52	0.79	0.40

Monologue 1	0.46	0.64	0.38
Monologue 2	0.6	0.80	0.56
Monologue 3	0.45	0.63	0.37

Realistic Kappa Group 1 – final validation

Agreement type	total	dl	mn
General agreement	0.65	0.66	0.63
Terminal breaks	0.81	0.80	0.80
Non terminal breaks	0.62	0.65	0.59

Group 2

First test:

Segmentation of an 822 word dialogue and a 1014 word monologue.

Results:

Group 2 Kappa – test 1

Agreement type	total	dl	mn
General agreement	0.75	0.78	0.73
Terminal breaks	0.77	0.81	0.73
Non-terminal breaks	0.62	0.58	0.63
No break	0.83	0.86	0.80

Group 2 Percentage – test 1

Agreement type	dl	mn
Total agreement	82%	78%
Terminal breaks	11%	6%
Non-terminal breaks	4%	9%
No break	67%	62%
Partial agreement	15%	20%
Terminal break vs. non-terminal	7%	7%
Non-terminal break vs. no break	8%	14%
Total disagreement	2.8%	2.2%
Terminal break vs. no break	1.6%	0.8%
Terminal break vs. non-terminal vs. no break	1.2%	1.4%

Second test:

Segmentation of a 1359 word dialogue.

Results:

Group 2 Kappa – test 2

Agreement type	dl
General agreement	0.76
Terminal break	0.82
Non-terminal break	0.57
No break	0.82

Group 2 Percentage – test 2

Agreement type	dl
Total agreement	79%
Terminal break	15%
Non-terminal break	5%
No break	59%
Partial agreement	17%
Terminal break vs. non-terminal	6%
Non-terminal break vs. no break	11%
Total disagreement	4.2%
Terminal break vs. no break	2.0%
Terminal break vs. non-terminal vs. no break	2.2%

Third test:

Segmentation of a 784 word monologue.

This test was taken by three out of four transcribers from group 2.

Results:

Group 2 Kappa – test 3

Agreement type	mn*
General agreement	0.68
Terminal breaks	0.78
Non-terminal breaks	0.51
No break	0.75

* 3 raters

Group 2 Percentage – test 3

Agreement type	mn*
Total agreement	79%
Terminal break	8%
Non-terminal break	7%
No break	64%
Partial agreement	20%
Terminal break vs. non-terminal	6%

Non-terminal break vs. no break	15%
Total disagreement	0.8%
Terminal break vs. no break	0.4%
Terminal break vs. non-terminal vs. no break	0.4%

* 3 raters

Fourth test:

Segmentation of a 681 word dialogue.

Results:

Group 2 Kappa – test 4

Agreement type	dl
General agreement	0.78
Terminal break	0.80
Non-terminal break	0.68
No break	0.84

Group 2 Percentage – test 4

Agreement type	dl
Total agreement	79%
Terminal break	11%
Non-terminal break	12%
No break	55%
Partial agreement	18%
Terminal break vs. non-terminal	7%
Non-terminal break vs. no break	11%
Total disagreement	3.2%
Terminal break vs. no break	1.5%
Terminal break vs. non-terminal vs. no break	1.8%

Fifth test:

Segmentation of an 803 word dialogue.

Results:

Group 2 Kappa – test 5

Agreement type	dl
General agreement	0.77
Terminal break	0.85
Non-terminal break	0.66
No break	0.81

Group 2 Percentage – test 5

Agreement type	mn
Total agreement	79%
Terminal break	12%
Non-terminal break	12%
No break	55%
Partial agreement	19%
Terminal break vs. non-terminal	5%
Non-terminal break vs. no break	14%
Total disagreement	2.4%
Terminal break vs. no break	0.7%
Terminal break vs. non-terminal vs. no break	1.6%

Sixth test:

Segmentation of an 1126 word monologue.

Results:

Group 2 Kappa – test 6 (3 raters)

Agreement type	mn
General agreement	0.77
Terminal breaks	0.82
Non-terminal breaks	0.66
No break	0.83

Group 2 Percentage – test 6

Agreement type	mn
Total agreement	80%
Terminal break	10%
Non-terminal break	9%
No break	61%
Partial agreement	19%
Terminal break vs. non-terminal	6%
Non-terminal break vs. no break	13%
Total disagreement	1.3%
Terminal break vs. no break	0.4%
Terminal break vs. non-terminal vs. no break	0.9%

Seventh test:

Segmentation of a 1045 word monologue.

Results:

Group 2 Kappa – test 7

Agreement type	mn
General agreement	0.79
Terminal break	0.76
Non-terminal break	0.70
No break	0.86

Group 2 Percentage – test 7

Agreement type	mn
Total agreement	84%
Terminal break	10%
Non-terminal break	12%
No break	63%
Partial break	14%
Terminal break vs. non-terminal	5%
Non-terminal break vs. no break	8%
Total disagreement	1.9%
Terminal break vs. no break	1.3%
Terminal break vs. non-terminal vs. no break	0.6%

Eighth test

Segmentation of a 981 word monologue.

Results:

Group 2 Kappa – test 8

Agreement type	mn
General agreement	0.82
Terminal break	0.83
Non-terminal break	0.75
No break	0.87

Group 2 Percentage – test 8

Agreement type	mn
Total agreement	87%
Terminal break	6%
Non-terminal break	8%
No break	73%
Partial agreement	12%

Terminal break vs. non-terminal	5%
Non-terminal break vs. no break	7%
Total	disagreement 1.5%
.....	
Terminal break vs. no break	0.7%
Terminal break vs. non-terminal vs. no break	0.8%

Realistic Kappa Group 2 – pre-validation

Text	Overall	Terminal	Non-terminal
Dialogue 1	0.51	0.67	0.45
Dialogue 2	0.52	0.71	0.47
Dialogue 3	0.57	0.74	0.52
Monologue 1	0.47	0.64	0.44
Monologue 2*	0.38	0.72	0.23
Monologue 3	0.54	0.75	0.50
Monologue 4	0.61	0.77	0.59
Monologue 5	0.54	0.68	0.51

*3 raters

Partial Kappa Group 2 – pre-validation

Text	Overall
Dialogue 1	0.86
Dialogue 2	0.82
Dialogue 3	0.84
Monologue 1	0.80
Monologue 2*	0.75
Monologue 3	0.83
Monologue 4	0.87
Monologue 5	0.86

* 3 raters

Validation: after the transcription and before the revisions

Test:

Segmentation of a 562 word dialogue and a 758 word monologue.

Results:

Final validation Kappa

Agreement type	total	dl	mn
General agreement	0.86	0.86	0.85
Terminal break	0.87	0.87	0.86

Non-terminal break	0.78	0.78	0.78
No break	0.91	0.91	0.90

Final validation Percentage

Agreement type	dl	mn
Total agreement	90.6%	90.7%
Terminal break	12.5%	8.3%
Non-terminal break	11.0%	12.4%
No break	67.1%	70.0%
Partial agreement	8.5%	9.1%
Terminal break vs. non-terminal	3.9%	3.7%
Non-terminal break vs. no break	4.6%	5.4%
Total disagreement	0.9%	0.2%
Terminal break vs. no break	0.7%	0.1%
Terminal break vs. non-terminal vs. no break	0.2%	0.1%

Realistic Kappa Group 1 – final validation

Agreement type	total	dl	mn
General agreement	0.65	0.66	0.63
Terminal breaks	0.81	0.80	0.80
Non terminal breaks	0.62	0.65	0.59

Partial agreement Kappa Group 1 – final validation

Agreement type	total	dl	mn
General agreement	0.91	0.91	0.90

Final results:

0.87 general agreement for terminal prosodic boundary annotation and 0.78 for non-terminal prosodic boundary annotation. Total general agreement for prosodic segmentation higher than 90%

The 95% confidence interval for prosodic segmentation convergence for the C-ORAL-BRASIL lies between 0.88 and 0.99.

Appendix 5 C-ORAL-BRASIL Transcription validation report

Content validated

- General mistakes:
 - Orthography mistakes, typing or transcription different from the form in the audio;
 - Word insertion (transcribed word, not found in the audio);
 - Word deletion (non-transcribed word present in the audio);
- Incorrect application of transcription criteria:
 - Word should have been transcribed with standard orthography but was transcribed following non-orthographic criteria;
 - Word should have been transcribed according to non-orthographic criteria but was transcribed following standard orthography;
 - Incorrect application of non-orthographic criteria, resulting in non-predicted form in both orthographic and non-orthographic criteria.

Checked utterances did not have overlappings. The examined samples were chosen randomly without data repositioning.

First validation (before the last revision)

Initial validation sample table

Sample	Size	Search
1A	7484 words (5% utterances from 89 texts)	All mistake types
1B	8949 words (10% utterances from 50 texts)	19 non-orthographic criteria

Results:

Sample 1A validation results. Search for all mistake types.

Mistake type	Initial validation - sample 1A	
	mistakes/words	%
All mistakes (1 + 2)	140/7484	1.87%
1. General mistakes (a + b + c)	104/6319	1.65%
a. Spelling mistake / inconsistency with audio	45/6319	0.71%
b. Word insertion	19/6319	0.30%
c. Word deletion	40/6319	0.63%
2. Mistake in non-orthographic criteria application	37/1165	3.18%

95% confidence interval for transcription accuracy at the end of the initial validation for the C-ORAL-BRASIL lies between 0.978 and 0.984.

In sample 1B there were 1,308 words to which non-orthographic criteria were applicable and there were only 8 mistakes, which correspond to an error percentage of 0.61%.

Final validation

Final validation sample table

2A	8243 words (5% utterances from 89 texts)	final	All mistake types
2B	8877 words (10% utterances from 50 texts)	final	19 non-orthographic criteria
2C	11805 words (5% utterances from de 139 texts)	final	3 non-orthographic criteria
2D	11215 words (5% utterances from 139 texts)	final	1 non-orthographic criterion

Results:

Sample 2A validation results. Search for all mistake types.

Mistake type	Final validation – sample 2A	
	mistakes/words	%
All mistakes (1 + 2)	67/8243	0.81%
1. General mistakes (a +b + c)	55/6124	0.90%
a. Spelling mistake / inconsistency with audio	23/6124	0.38%
b. Word insertion	19/6124	0.31%
c. Word deletion	13/6124	0.21%
2. Non-orthographic criteria application mistake	12/2119	0.57%

The 95% confidence interval for transcription accuracy of the C-ORAL-BRASIL after the final validation lies within 0.989 and 0.993.

In sample 2B there were 693 words to which non-orthographic criteria were applicable and there were only 3 mistakes found, therefore the error percentage was 0.43%.

Sample 2C was probed for 3 non-orthographic criteria (apheresis, reduced prepositions, first person plural pronoun). There were 66 words to which the criteria were applicable (40 apheretic forms, 12 prepositional forms, 14 first person plural pronominal forms). Only one transcription mistake was found in reference to a prepositional form transcribed according to non-orthographic criteria whereas it conformed to standard orthography. The error percentage in sample 2C was 1,5%.

Sample 2D was probed for preposition spelling. There were 16 prepositions transcribed in reduced form in this sample (c', p', pr', d', n') and 23 transcribed in standard orthographic form (com, para, de, em). No mistake was found in the transcription of the examined data (39 prepositions overall).