Challenges in modality annotation in a Brazilian Portuguese Spontaneous Speech Corpus

Luciana Beatriz Avila
PosLin-UFMG/UFV/Capes
Av Antonio Carlos 6627
Belo Horizonte, MG 31270-901 Brazil
lucianabeatrizavila@gmail.com

Heliana Mello
UFMG/FGV/CNPq
Av Antonio Carlos 6627
Belo Horizonte, MG 31270-901 Brazil
heliana.mello@gmail.com

Abstract

This short paper introduces the first notes about a modality annotation system that is under development for a spontaneous speech Brazilian Portuguese corpus (C-ORAL-BRASIL). We indicate our methodological decisions, the points which seem to be well resolved and two issues for further discussion and investigation.

1 Credits

The authors are thankful to CNPq, FAPEMIG and CAPES (Proc. nº BEX 9537/12-0) for research funding support.

2 Introduction

Modality annotation is inexistent for both written and spoken Brazilian Portuguese corpora, thus the novelty of this project. According to Nurmi (2007:1), linguistic annotation is helpful for the recovering of linguistic elements; however, the multifaceted nature of modality "is still a hurdle in computer assisted-research". Following up on the same reasoning, Baker et al. (2010: 1403) say that "[t]he challenge of creating a modality annotation scheme was to deal with the complex scoping of modalities with each other and with negation, while, at the same time creating a simplified operational procedure that could be followed by language experts without special training". Therefore, understanding what this semantic category stands for, as well as identifying linguistic elements that carry it, is of utmost relevance.

Our goal in annotating modality in a spontaneous speech Brazilian Portuguese Corpus is to provide a reliable starting point for researchers that might be interested in developing methodologies associated to NLP that ensue the extraction of oral discourse reliability, certainty and factuality markers, or carrying sentiment analysis, modeling modality and similar objectives.

3 Defining modality

In this paper we study modality in a spontaneous speech corpus, the C-ORAL-BRASIL, which will be presented in 4 below. As for spontaneous speech, we follow Cresti and Scarano (1998:5) in characterizing it as "the fulfillment of linguistic acts, not programmed and not programmable, because they emerge during the unfolding of an interaction, always new and unpredictable, between interlocutors." Their view is based on Austin's (1962) Speech Act Theory that associates spoken text to the realization of speech acts. According to Cresti and Scarano spontaneous speech is governed by an illocutionary principle, not found in written texts, as well as specific informational articulations.

Modality is taken here in Ballinian terms, that is, it stands for the evaluation or the point of view of a subject who evaluates the locutory material in a given utterance in a communicative act. Since the domain of our analysis is the spoken text, we follow Cresti's (2000) Language into Act Theory, whereby the utterance is the analytical reference unit that will be taken into consideration. This significantly differs from studies that rely on the sentence as the reference unit for the analysis of modality (eg. Fintel, 2006). An utterance carries an illocution and its locutory material does not necessarily carry a proposition. An utterance may be simple when comprised by one tone unit or complex when it is made up by two or more tone units. The scope of modality is the tone unit as proposed by Tucci (2007). Hence, within a given complex utterance, there might be different tone units which carry different modal values. When a tone unit carries more than one modal marker they may not share the same modal value, in which case the dominant modality will prevail. This can be appreciated in the examples below:

(i) REN: **se** a gente vai de táxi / voltar de táxi / **po'** comprar um //
If we go by taxi / come back by taxi / (you) may buy one //

In (i) there are three information units compounding a complex utterance. The first one carries epistemic modality while the last one carries deontic modality. Albeit belonging to the same utterance, the modalities that mark each information unit are not semantically compositional. Whereas in (ii) below, two modal values within the same information units will be compositional and the dominant value will prevail:

(ii) GIL:<eu **acho** que **tem que** ser esses> //
I think that it has to be these

In (ii) there is a single information unit, a Comment, which carries two modality indexes, *acho* "think", an epistemic marker, and *tem que* "have to", a deontic marker. The utterance in this case carries only one information unit and its modality is epistemic.

Modality in speech at times might get confused with other categories that carry subjective judgment; however a good rule of thumb to identify modal markers is to proceed to a semantic analysis leaving strictly pragmatic values aside. This has been demonstrated through an experiment reported in Mello and Raso (2011) who indicate the differentiation between modality, illocution and attitude in speech**.** Modality is related to the modus on dictum (Bally, 1942), illocution is the actum of the dictum, while attitude is the modus on actum. Therefore, modality can be classified as a semantic category, whereas both illocution and attitude are pragmatic notions. Modality, when marked, is carried by lexical and grammatical items on the one hand, while illocution and attitude are carried by prosodic cues.

In our work we focused on overt modal markers and took into consideration the following modal values: epistemic, deontic and dynamic. Epistemic modality refers to the conceptualizer's point of view, as far as possibility and necessity are concerned, in a given uttered material. This can be seen in the example below:

(iii) REN: **<**pode**> // tanto faz //** pode **//**
<It can be> // it doesn't matter // it can be //
FLA: ou cê **acha** muito **//**
Or do you think this is too much //
REN: uhn **//** acho **que não //**
Uhn // I don't think so //

In deontic modality the conceptualizer, a moral agent, refers to obligation, permission, contingent necessity in the uttered locutory material, as in (iv):

(iv) HMB: ela **tem que** falar / assim / de que que ela gosta //
She has to say / like/ what she likes //

As for dynamic modality, it includes ability and intention (will), that is, the conceptualizer's expression of capability, as in (v):

(v) ROG: eu acho que eu **consigo** &mar [/1] fazer isso //
I think I can do this //

Modality is usually codified by several morphological and grammatical forms, among them modal auxiliaries, adverbs, evaluative adjectives, periphrastic forms, propositional verbs and conditionals. These forms will be taken into consideration in the proposed annotation system whereas some less conventionalised items that might be becoming grammaticalized in spoken Brazilian Portuguese will not be annotated because they require further investigation.

Our annotation proposal is inspired by other systems previously explored for English (Baker et al., 2010; Matsuyoshi et al., 2010; Saurí et al., 2006, 2007; Szarvas et al., 2008) and it closely follows the scheme proposed by Hendrickx, Mendes and Mencarelli (2012) who explored modality annotation in European Portuguese (EP) speech.

The EP proposed scheme takes into account seven modal values and a number of corresponding subvalues, as shown in Table 1:

Values	**Subvalues**
Epistemic	knowledge belief doubt possibility interrogative
Deontic	obligation permission
Participant-internal	necessity: personal needs capacity: personal capacity
Evaluation	(evaluation of the proposition)
Volition	(hopes and wishes)
Effort	(attempt of the participant to make sth. happen)
Success	(results of the commitment of the participant)

Table 1 – European Portuguese selected modal values and subvalues

The system we advance here is more economical and reflects a canonical typology of modal meanings, as we show below. In both schemes (EP and BP), the three main categories overlap (Epistemic, Deontic, Dynamic or Participant-Internal), but it is not sufficiently clear so far whether a variety of non-epistemic meanings taken into consideration in the EP system should be considered as separate modal values, or rather as subvalues of Epistemic modality.

Other related works on modality annotation, accordingly to their goals, also present a range of modal values, denoting requirement, permissiveness, intention, ability, effort, success want and belief (Baker et al. 2010); assertion, volition, wish, imperative, permission, interrogative (Matsuyoshi et al. 2010); purpose, need, obligation and desire events (Morante & Daelemans, 2012).

Much of these works describes other components which are involved in the expression of modality, such as trigger, target and holder (Baker et al. 2010) or source, time, conditional, primary modality type, actuality, evaluation and focus (Matsuyoshi et al. 2010).

Following on the footsteps of the annotation scheme for EP, our proposal aims at contributing to the development of NLP projects, especially those based on spontaneous speech and its particularities.

4 A Brazilian Portuguese spontaneous speech corpus: the C-ORAL-BRASIL

C-ORAL-BRASIL follows the same architecture as the European Romance spontaneous spoken corpus C-ORAL-ROM (Cresti and Moneglia, 2005), whereby diaphasic variation is privileged in order for a large diversity of illocutions and informational structuring to be documented. C-ORAL-BRASIL comprises 200 texts of approximately 1,500 words each. Its informal half has been published (Raso and Mello, 2012) and exhibits a majority of private/familiar texts (80%) over public texts (20%), equally distributed into dialogues, conversations (3 or more participants) and monologues. The corpus follows the CHILDES-CLAN transcription format to which prosodic annotation is added, marking tone unit and utterance boundaries, besides several phenomena typical of speech. The entire corpus is speech to text aligned through use of WinPitch software.

5 Annotating modality in the C-ORAL-BRASIL

In this study a sample from the C-ORAL-BRASIL was taken into consideration. It covers 20 texts with an average of 1,500 words each, thereby totally 31,318 words; 5,484 utterances and 9,825 tone units. 1,155 modality marked tone units were found. The identification of modal markers was undertaken by three annotators working independently and qualitatively validated through group discussions. The search for modal markers was performed manually, through qualitative textual analysis, supported by the WinPitch software which allows for the concomitant examination of speech signal and transcription.

The data were organized in a table containing the modal markers, the tone unit in which they occurred, the type of information unit they are inserted in, the file they belong to, and any qualitative information deemed relevant.

The modality annotation scheme we propose takes into account The Language Into Act Theory and its reference unit, the utterance, and its subunits, that is, information units (Cresti, 2000). The scope of modality also follows the proposal established within that theory, thereby assigning its locus to the information unit (Tucci, 2007). Additionally, as mentioned, previous work on English and European Portuguese modality annotation is observed closely (cf. Section 3) in addition to opinion and emotions annotation (Wiebe et al., 2005).

The methodological steps taken in order for us to arrive at a modality annotation system were the following: the listing of a set of modal values emerging from the modal indexes found in the corpus; these values were subsequently tested on a sample of our subcorpus.

For the purpose of modality annotation we consider three modal values: epistemic, deontic and dynamic. As discussed above, epistemic modality relates to the conceptualizer's commitment to a given locutory material. Epistemic modality carries seven subvalues: knowledge, opinion, belief, possibility, probability, necessity and verification. [1] Deontic modality encompasses four subvalues: obligation, permission, prohibition and restriction. Finally, dynamic modality comprises three subvalues: ability, capacity and volition/intention.

In addition to modal values, the annotation scheme is made up of the following elements:

- **Trigger (M)**: the morpholexical and grammatical items that carry modality;
- **Source of the modality (src_mod)**: the conceptualizer, who might coincide with the speaker , the addressee, or another individual whose perspective and view point is being reported;
- **Source of the event mention (src_evt):** the producer**,** the speaker;
- **Target (T)**: the expression in the scope of the trigger within an annotation unit, that is, information units (IU)that carry modality (Comment, Topic, Parenthetical, Locutive Introducer), described in Table 2:

	IU	**Information function**
Textual units	Comment	Expresses the illocutionary force of the utterance
	Topic	Specifies the *locus* of application of the illocutionary force of the Comment
	Parenthetical	Expresses metalinguistic integration of the utterance
	Locutive Introducer	Signals pragmatic suspension of the *hic et nunc* and introduces a metaillocution

Table 2 – Modalized textual units (Cresti, 2000)[2]

An example of a modality annotated utterance is given below. Due to space constrains, we cannot discuss all the details involved in the process, however, it is relevant to note the following: elements within = marks stand for information unit labeling, angled brackets stand for speech overlaps, square brackets stand for modality annotated elements, single slashes stand for non-terminal breaks and double slashes for terminal breaks.

(vi) EVN$_{\mathbf{S1}}$: é / [a <gente]$_{\mathbf{S1}}$ [tem que]$_{\mathbf{M}}$> <[restringir também]$_{\mathbf{T}}$ / isso> //
Yeah / we have to restrict too / this //

The annotated elements are the following:

Trigger	tem que
Source of modality	A gente, 1p
Source of event mention	EVN
Modal value	deontic_obligation
Target	restringir também

Example (vi) is very straightforward and leaves no room for discussion as far as modality labeling

[1] We include here factuality and certainty features.

[2] Adapted from Tucci (2007).

and domains are concerned. However, this is not all we see in the data analyzed. There is plenty for discussion regarding some complex issues. Two of these are briefly mentioned below.

One of the challenges is the characterization of the elements that fulfill the role of Source. In our sample we found a majority of cases in which the conceptualizer overlaps with the first person speaker (cf. vi). However, there are cases in which the speaker presupposes or evaluates the kind of modal judgment that is made by others, in which case apparently there could be two conceptualizers, whereby two Source roles would be assigned, S_1 and S_2 and the assigned modal value would be shared by them (cf. vii). Yet another case occurs when the speaker reports the modal judgment made by a third party, in which case, the speaker does not partake in the modal conceptualization that is overtly manifested (cf. viii).

(vii) **JOR**$_{\mathbf{S1}}$: se o brasileiro nũ lê os manuais /=TOP= hhh no mercado de reposição / &auto [/1] de autopeça / **eles**$_{\mathbf{S2}}$ acham que abrir uma empresa é comprar um produto por um real / na base cem / e vender por dois acha que tá ganhando o &do [/2] o dobro //

If Brazilians don't read manuals / ..../ **they** think that to open a business is to buy a product by one real/ …/ and to sell it for two (they) think they are making double //

(viii) **PAU**: e **a Isa**$_{\mathbf{S1}}$ tava achando que ela ali ia ficar pequena //

And **Isa** was thinking that it would be small //

These two last examples lead us to mark up two different sources, following the annotation scheme proposed by Hendrickx, Mendes and Mencarelli (2012): Source of the event mentioned and Source of the modality. The first one corresponds to the producer of the sentence with the modal marker; the second one to the person who is agent/experiencer of modality.

As pointed out by Saurí et al.'s FactBank annotation scheme (Saurí, 2008; Saurí and Pustejovsky, 2009), there is always a default source[3] corresponding to the author of the text and "the factuality value assigned to events in text must be relative to the relevant sources at play in the discourse […]" (Saurí and Pustejovsky, 2009, p. 240).

In (vii), we have at least two different event mentions[4] introduced by modal markers (e_1, by "se", and e_2, by the epistemic verb "achar"). The difference between event e_1 and event e_2 is that in e_1 the Source of the modality and the Source of the event mentioned overlap ("JOR") and refer to the epistemic judgment expressed by the conditional construction, whereas in e_2 Source of the event mentioned ("JOR") and Source of the modality ("eles") are different entities. The relation between S1 and the epistemic judgment of S2 is based on a supposition on the evaluation of the second conceptualizer, not necessarily corresponding to the truth-value of the uttered material.

Finally, in (viii), Source of the modality ("a Isa") and Source of the event mentioned ("PAU") are explicitly distinct. There is just the third-person conceptualizer, "a Isa", and her epistemic judgment is reported by "PAU".

A second challenge is presented by the labeling of target. In default circumstances, the target shares the same information unit as the modal marker, as posited by the Language into Act Theory, and which can be seen in the examples previously presented. However, there are cases in which there seems to be a percolation of the target through information unit boundaries, as if it were an anaphoric element, as can be seen in (ix) below:

(ix) GIL: <ô / mas> / voltando à questão / falando em e também falando em povo mascarado / esse povo do Galáticos é muito palha / eu **[acho que]**$_{\mathbf{M}}$ [es nũ deviam mais participar / e <tal>]$_{\mathbf{T}}$ //

(...) / I think that they shouldn't participate anymore / like //

LEO: <**[com certeza]**>$_{\mathbf{M}}$ //
Certainly //

[3] "Sources are understood here as the cognitive individuals presented as holding a specific stance with regards to the factuality status of events in text." (Saurí, 2008, p. 58).

[4] An event mention is defined as "consisting of a core predicate and its arguments (complements and adjuncts) in the sentence." (Matsuyoshi et al., 2010, p. 1458).

In the above example, “com certeza” refers back to the deontic assertion made in the previous turn “es nũ deviam mais participar”; however it is not clear how this can be annotated within the present scheme. One possible solution could be to add a Comment slot, in which we annotate the anaphoric reference.

6 Final remarks

In this short paper we have introduced the first notes about a modality annotation system that is being developed for a Brazilian Portuguese spontaneous speech corpus. Although we were able to point to some efficient methodological solutions we have implement so far, much remains open for discussion and further investigation.

.

References

Baker, K., Bloodgood, M., Dorr, B. J., Filardo, N. W., Levin, L., Piatko, C. 2010. A modality lexicon and its use in automatic tagging, in Proceedings of the Seventh Language Resources and Evaluation Conference (LREC’10).

CHILDES - Child Language Data Exchange System. Available at: http://childes.psy.cmu.edu/ Accessed: 29 Nov. 2012.

Cresti, E. 2000. *Corpus di italiano parlato*. Firenze: Accademia della Crusca.

Cresti, E.; Scarano, A. Sur la notion de parlé spontané. Available at: http://lablita.dit.unifi.it/preprint/preprint-98coll03.pdf. Accessed: 15 Nov. 2012.

Cresti, E.; Moneglia, M. 2005. *C-ORAL ROM*: Integrated reference corpora for spoken Romance languages. Amsterdam/Philadelphia: John Benjamins.

Hendrickx, I.; Mendes, A.; Mencarelli, S. Modality in Text: a Proposal for Corpus Annotation. in Proceedings of the Eighth conference on the International Language Resources and Evaluation (LREC’12), pp.1805-1812.

von Fintel, K. Modality and Language. In Encyclopedia of Philosophy – Second Edition, edited by Donald M. Borchert. Detroit: MacMillan Reference USA.

Martin, P. WinPitch Corpus. Available at: http://www.winpitch.com. Accessed: 15 Nov. 2012.

Matsuyoshi, S., Eguchi, M., Sao, C., Murakami, K., Inui, K., Matsumoto, Y. 2010, Annotating Event Mentions in Text with Modality, Focus, and Source Information, in Proceedings of the Seventh conference on the International Language Resources and Evaluation (LREC’10).

Mello, H. R.; Raso, T. 2011. Illocution, modality, attitude: different names for different categories. In: Mello, H. R.; Panunzi, A.; Raso, T (eds.). *Pragmatics and prosody*: illocution, modality, attitude, information patterning and speech annotation. Firenze: Firenze University Press, pp. 1-18.

Morante, R.; Daelemans, W. 2012. Annotating modality and negation for a machine reading evaluation. In: P. Forner, J. Karlgren, & C. Womser-Hacker (eds.), CLEF 2012 Conference and Labs of the Evaluation Forum - Question Answering For Machine Reading Evaluation (QA4MRE), Rome, Italy.

Nurmi, A. Employing and elaborating annotation for the study of modality. Available at: www.helsinki.fi/varieng/journals/volumes/01/nurmi/ . Accessed: 27 Mar. 2012.

Raso, T.; Mello, H. 2012. *C-ORAL-BRASIL I*: Corpus de referência do português brasileiro falado informal e DVD multimedia. Belo Horizonte: Editora UFMG,. v. 1.

Saurí, R., Verhagen, M., Pustejovsky, J. 2006. Annotating and recognizing event modality in text, in Proceedings of the 19th International FLAIRS Conference, FLAIRS 2006.

Saurí, R. 2008. A Factuality Profiler for Eventualities in Text. Ph.D. Thesis. Brandeis University.

Saurí, R., J. Pustejovsky. 2009. FactBank: A Corpus Annotated with Event Factuality. Language Resources and Evaluation. Available at: http://link.springer.com/article/10.1007/s10579-009-9089-9/fulltext.html. Accessed: 28 jan. 2013.

Szarvas, G., Vincze, V., Farkas, R., Csirik, J. 2008. The BioScope corpus: annotation for negation, uncertainty, and their scope in biomedical texts, in Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing, Association for Computational Linguistics, pp. 38-45.

Tucci, I. 2007. La modalità nel parlato spontaneo e il suo dominio de pertinenza. Una ricerca corpus-based (C-ORAL-ROM italiano). Actes du XXVe CILPR. (Innsbruk 3-8 September 2007). Available at: http://lablita.dit.unifi.it/preprint/preprint.2008-02-06.3658867998 Accessed: 13 jan. 2013.

Wiebe, J., Wilson, T., Cardie, C. 2005. Annotating expressions of opinions and emotions in language. Kluwer Academic Publishers, pp. 1-54.