Modality in Spoken Texts: a Proposal for Corpus Annotation

Silvia Mencarelli, **Iris Hendrickx** and **Amália Mendes**

This paper presents an annotation scheme of modality developed for Portuguese, and its application to the spoken corpus of European Portuguese C-ORAL-ROM (Bacelar do Nascimento et al., 2005).

Modality is considered an extra-propositional component of meaning (Baker et al., 2010), as it is related to the way meaning is presented and interpreted, and is usually defined as the expression of the speaker's opinion and of his attitude towards what he is saying (Oliveira, 2003; Palmer, 1986).

The annotation of modality for Portuguese is still an unexplored area both for spoken and written corpora. The values included in the annotation scheme that we propose in this paper are based on the Linguistics literature for English (Palmer, 1986; van der Auwera et al., 1998) and for Portuguese (Oliveira, 1988) and on their comparison to the ones identified in the literature on the annotation of modality (Baker et al., 2010; Matsuyoshi et al., 2010; Saurí et al., 2006; Wiebe et al., 2005). Most of the existing research on the annotation of modality, however, is not only centered on the identification of modalities, but also includes other types of semantic information, as a result of a more general purpose of improving Natural Language Processing applications.

The annotation scheme for modality we propose has four components:

- trigger: the lexical element conveying the modal value;
- source: we identify two kinds of source, the source of the event mention (speaker) and the source of the modality (agent or experiencer);
- target: the expression in the scope of the trigger;
- modal value.

In Example 1 we apply our annotation scheme to a sentence extracted from the C-ORAL-ROM corpus:

(1) *A: para saia e casaco **tem de** ser cano alto //$*

Source: speaker (A)

Trigger: tem de

Target: ser cano alto

Modal value: deontic_obligation

The modal values and subvalues we identify are:

- epistemic modality to annotate the commitment of the speaker towards the truth of the proposition. We identify six subvalues, like epistemic_possibility to annotate when the speaker expresses a weak belief in the truth of the proposition, and epistemic_necessity when the speaker expresses a strong belief (the other subvalues are epistemic_knowledge, epistemic_belief, epistemic_doubt and epistemic_interrogative).
- participant-internal modality, focusing on the conditions which make the speaker engage in the state of affairs, to annotate when the speaker expresses a personal need (participant-internal_necessity) or capacity (participant-internal_capacity);
- deontic modality to annotate when the speaker imposes something on the hearer: we mainly identify deontic_obligation (for commands) and deontic_permission (for permissions) (the other subvalues are deontic_suggestion and deontic_request);
- commissives to annotate the speaker's commitment to do something (e.g. promises and threats);
- volition, for hopes, whishes or desires;
- evaluation, for the speaker's evaluation on the proposition.

We also observe two modalities proposed by Baker et al. (2010):

- effort, when the source tries to achieve a goal;
- success, when the verb expresses if the participant succeeded or not in his objective.

A preliminary study of the C-ORAL-ROM corpus showed us that the same type of modal expressions are used in both spoken and written data, although their frequency can

show some variation. However, we do notice that we need to consider additional typical elements of spoken corpora (e.g. discourse markers, repetitions, pauses and intonation).

Some elements carrying modal information are:

- a pause, silent or filled by an extra-linguistic marker, may indicate some doubt about what the speaker says and thus have an epistemic meaning:

  (2) *A: e / o traje académico / não é muito usado?$*
  *B: &**hum** / depende //$*

  (3) *A: no conjunto / está-se melhor aqui //$ ainda que / a vida seja mais cara //$ mas também / há mais possibilidades de se ganhar mais //$*
  *B: **hum** //$ quando / quando tu dizes / vive-se mais / o que é que queres dizer?$*

- some typical oral expressions such as *vamos lá ver* or *ora bem* may express epistemic doubt, when produced after something said by another speaker (4) or at the beginning of an hypothesis (5), and epistemic belief (6);

  (4) *A: eu acho que o cortar com a língua francesa / implica uma viragem também nesse continuum <cultural> //$*
  *B: [<] <pois> //$ **vamos lá ver** //$ porque / o francês já está / desde há uns anos para cá / numa situação de inferioridade em relação ao inglês //$*

  (5) ***Vamos lá ver** se somos capazes de dizer isto com rapidez //$*

  (6) *A: **ora bem** / &ah / esta é uma exposição / de / digamos / um bocadinho ligada aquela preocupação / de mostrar / os trabalhos do cinema de animação português /*

- in spoken data, the listener uses expressions like *não acredito, a sério?, não pode*, to show epistemic doubt or surprise regarding what the speaker says (7);

  (7) *A: vinha no jornal / que eles tinham assaltado os carros no parque de estacionamento num concerto / em setembro / daquele ano //$*
  B: [<] <&ah / que engraçado> //$ não tinha feito a ligação //$
  A: sim //$ telefonei para o jornal e tudo //$
  C: <hhh>$

B: / [<] <não te lembras> ?$

A: [<] ***<estás a brincar>?$***

B: não //$ <telefonei para o jornal> //$

A: [<] ***<estás a falar a sério>?$***

- expressions as *sim*, *eu sei* (8), *é é* (9), *pois* (10), marking agreement or understanding, may be related to epistemic knowledge;

(8) *A: olha / &eh / daqui mãe //$*

B: ***sim*** */* ***eu sei*** */ já <percebi / hhh> //$;*

(9) *A: / [<] <é uma questão> de insegurança //$*

*B: é //$ <****é*** */* ***é****> //$*

(10) *A: [<] <mas isso depende> de como as vais usar //$ com saias tem que ser altas //$*

*B: <****pois****> //$*

In the final version of this paper we present the annotation of approximately 100 utterances, and report in more detail on the differences in modality between spoken and written language.

BIBLIOGRAPHY

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).

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 International Language Resources and Evaluation (LREC'10).

OLIVEIRA, F., (1988), *Para uma semântica e pragmática de DEVER e PODER*, Dissertação de Doutoramento em Linguística Portuguesa apresentada à Universidade do Porto, Faculdade de Letras.

OLIVEIRA, F., (2003), *Modalidade e modo*, in *Gramática da Língua Portuguesa*, Lisboa, Editorial Caminho, pp. 243-272.

PALMER, F. R., (1986), *Mood and Modality*, Cambridge textbooks in linguistics, Cambridge University Press, Cambridge.

SAURÍ, R., VERHAGEN, M., PUSTEJOVSKY, J., (2006), *Annotating and Recognizing Event Modality in Text*, in Proceedings of the 19th International FLAIRS Conference, FLAIRS 2006.

VAN DER AUWERA, J., PLUNGIAN, V., (1998), *Modality's semantic map*, in *Linguistic Typology 2*, pp. 79-124.

WIEBE, J., WILSON, T., CARDIE, C., (2005), *Annotating Expressions of Opinions an Emotions in Language*, in *Kluwer Academic Publishers*, pp. 1-54.

Reference Corpus:
BACELAR DO NASCIMENTO, M. F., BETTENCOURT GONÇALVES, J., VELOSO, R., ANTUNES, S., BARRETO, F., AMARO, R., (2005) *The Portuguese Corpus*, Chapter 3, in Cresti, E., Moneglia, M. (editors) , *C-ORAL-ROM, Integrated Reference Corpora fo Spoken Romance Languages*, John Benjamins Publishing Company, pp. 163-207.