Modality in Text: a Proposal for Corpus Annotation

Iris Hendrickx, Amália Mendes and Silvia Mencarelli

Centro de Linguística da Universidade de Lisboa
Av. Prof. Gama Pinto, 2
1649-003 Lisboa - Portugal
iris@clul.ul.pt, amalia.mendes@clul.ul.pt, tittibum@libero.it

Abstract

We present a annotation scheme for modality in Portuguese. In our annotation scheme we have tried to combine a more theoretical linguistic viewpoint with a practical annotation scheme that will also be useful for NLP research but is not geared towards one specific application. Our notion of modality focuses on the attitude and opinion of the speaker, or of the subject of the sentence. We validated the annotation scheme on a corpus sample of approximately 2000 sentences that we fully annotated with modal information using the MMAX2 annotation tool to produce XML annotation. We discuss our main findings and give attention to the difficult cases that we encountered as they illustrate the complexity of modality and its interactions with other elements in the text.

Keywords: Modality, Corpus Annotation, Portuguese

1. Introduction

This paper presents a scheme for the annotation of modality in Portuguese. Modality is usually defined as the expression of the speaker's opinion and of his attitude towards what he is saying (Palmer, 1986). In Portuguese, it has been studied from the theoretical linguistic perspective (especially (Oliveira, 1988)). However, a concrete proposal for the annotation of modality in Portuguese does not yet exist. Moreover, practical annotation schemes for languages other than English are a rare good.

In recent years we see a clear trend in information extraction applications to go beyond the extraction of pure facts, to focus on personal opinions in sentiment analysis and opinion mining (Wiebe et al., 2005), and to distinguish between factual an probable information (Saurí et al., 2006), to detect uncertainty, speculation and negation in biomedical text mining (Szarvaz et al., 2008; Baker et al., 2010; Matsuyoshi et al., 2010). This interest has resulted in some proposals for the annotation of modality mainly for the English language. Our annotation scheme for Portuguese is therefore based on the modal typologies presented in the theoretical literature on modality in English (Palmer, 1986) and Portuguese (Oliveira, 1988) and on the tangible annotation schemes proposed for English designed from a practical NLP viewpoint.

This paper is structured as follows. We first present related work. In section 3. we present our annotation scheme, listing the modal values we identify and the other components involved in the expression of modality. We also present annotation details such as language related problems and we discuss the annotation software that we used. In section 4. we describe the corpus sample of approximately 2000 sentences fully annotated with modal values. We discuss our main findings in 5. and several encountered difficult cases including some cases where the proposed scheme does not capture all depths of the phenomenon of modality in language in section 6. We conclude in section 7. and present our plans for future work.

2. Related work

The literature on modality proposes different typologies. In linguistics, most modal systems are based on the contrast between epistemic and deontic modality. The first is related to the notions of knowledge and belief, and points out the degree of commitment to the truth of the proposition, while the second deals with the notions of command and permission. While the epistemic value is stable across typologies, the other values that are contrasted with epistemic modality vary considerably. Some proposals distinguish between epistemic, participant-internal and participant-external modality (Van der Auwera and Plungian, 1998), or between epistemic, speaker-oriented modality and agent-oriented modality (Bybee et al., 1994). Other values generally considered are, for example, volition, related to the notions of will, hope and wish; evaluation, concerning the speaker's evaluation of facts and situations; and commissives, used by the speaker to express his commitment to make something happen (Palmer, 1986). In the literature on practical corpus annotation of modality, the attention focuses on the distinction between factual and non-factual information, as many NLP applications need to know what is presented as factual and certain and what is presented as non-factual or probable. This is related to epistemic modality, although it does in fact cover a larger number of linguistic contexts. Most annotation schemes also include orders and permissions (deontic) and wishes and wants (volition) (e.g. (Baker et al., 2010)). Opposed to the theoretical typologies of modality, these schemes also describe in detail which elements in the text are actually involved in the expression of modality and what roles do they have. These are the subject of the modality (source) and the elements in the scope of the modality (target/scope/focus). Other schemes ((Baker et al., 2010; Matsuyoshi et al., 2010; Saurí et al., 2006) also determine the relation between sentences in text, identifying temporal and conditional relations between events or the evaluation of the degree of relevance of some information within a text, rather than classifying modal values.

Not only the intention and goal of the researchers, but also

their background determines how and what will be annotated in an annotation scheme. This fact is demonstrated in the article of Vincze et al (Vincze et al., 2011), which analyzes how the same data, MEDLINE abstracts, was annotated in two different projects with the same kind of information, negation and uncertainty, but with completely different results and not much overlap. In our annotation scheme we have tried to combine a more theoretical linguistic viewpoint with a practical annotation scheme that will also be useful for NLP research but is not geared towards one specific application. Our notion of modality focuses on the attitude and opinion of the speaker or of the subject of the sentence.

Our approach is very similar to the approach taken in the OntoSem (McShane et al., 2005) annotation scheme for modality (Nirenburg and McShane, 2008). In the OntoSem scheme, phrases or words that trigger modality are labeled with a modal value and (here our annotation differs) a weight expressing the degree of modality. Beside the trigger, the scheme also defines a source ("attributed-to") and target ("scope") of the modality. They distinguish ten different modal values that closely overlap with the values in our scheme but there are also some differences. For example, the OntoSem modality scheme has a value "intentional" to capture expressed goals like in the phrase "we aim to" which we included as effort (which implies intention). In our scheme, we have the value doubt as a separate sub value of epistemic while they label these cases all as "belief". The modality annotation in OntoSem is part of a large and detailed semantic annotation scheme that also captures other types of semantic relations like semantic roles. On the contrary we only try to describe modal relations separately from any further syntactic or semantic information.

3. Annotation scheme

Here we introduce our annotation scheme for modality in Portuguese in which we combine a practical annotation with a theoretically-oriented perspective which focuses on a detailed variety of modal values and not on components such as factuality or conceptual relations. The final annotation scheme was created in several steps. On the basis of existing literature discussed in section 2. we first created an initial scheme of modal values and components that we would like to annotate, and made a first draft of the annotation guidelines. Next we applied the scheme to a corpus sample of sentences containing a potential modal verb (discussed in more detail in the next section) to evaluate whether our typology of modal values had the right level of detail and whether we covered all types of modality that were present in actual data. Annotation was done by one annotator and all difficult cases were discussed with a second annotator. We made several adaptations to the initial scheme, for example, at first we had a separate modal value for commissives (Palmer, 1986), but during annotation it turned out to be low frequent and most of the times overlapping with deontic obligation so we decided not to keep this as separate value. We also kept record of difficult cases and added some examples to the annotation guidelines. Here we present the final version of the annotation scheme.

First of all, we only annotate modal events and not entities. We consider eight main modal values and several sub values. Epistemic modality denotes the commitment of the speaker (or the participant, usually the subject) towards the truth of the proposition. We identify five sub-values: epistemic knowledge to annotate when the speaker presents his or someone else's knowledge or when he expresses some degree of understanding about something; epistemic belief, epistemic doubt, epistemic possibility, to annotate when the speaker presents what he or someone else is saying as a belief, a possibility or a probability and epistemic interrogative to denote questions. Deontic modality denotes when the speaker imposes something on the hearer. We identify two sub values: deontic obligation and deontic permission. As mentioned above, we also consider contexts with commissive value as a subtype of deontic obligation, since the speaker or participant establishes an obligation upon himself, as in example 1[1].

(1) Dirigente do Bloco de Esquerda *promete* apoiar luta anti-aterro na Figueira da Foz. 'The leader of the Bloco de Esquerda promised to support the fight against landfill utility at Figueira da Foz'.

Contrary to Van der Auwera and Plungian (1998), we don't consider participant-external modality as an independent type, but rather as a subtype of deontic modality. This decision is mostly due to the difficulty in establishing whether the obligation or the permission (possibility) is the responsibility of an animated entity or of some external conditions not controlled by the participants of the event. For example, in 2 it is difficult to define what or who establishes the necessity.

(2) Desse ponto de vista penso que é *preciso* avançar. 'From that point of view I believe we must go on'.

However we did follow the work of Van der Auwera and Plungian (1998) in marking up participant-internal necessity, to tag personal needs of the speaker or participant, and participant-internal capacity, to tag personal capacities of the speaker or participant, as in example 3 where the verb *saber* 'know' expresses the capacity of the player to play.

(3) Está permanentemente a pensar no cesto e *sabe* jogar muito bem. 'He is always thinking about the game and can play really well.'

We use the value evaluation to annotate the speaker's or participant's evaluation of the proposition, and the value volition for hopes and wishes. Following Baker et al. (2010) we consider the modalities effort and success: the first focuses on the attempt of the participant to make something happen, being expressed by *esforçar* 'put effort' in 4, and the latter focuses on the results of the commitment of the participant, being expressed by *conseguir* 'succeed' in 4.

(4) *Esforçou-se* tanto para aprender a nadar mariposa que em menos de um ano *conseguiu*. '[She] worked so hard to learn how to swim the butterfly stroke that in less than a year she succeeded'.

[1] All examples in this paper are taken from the corpus sample discussed in section 4.

Besides identifying modal values, we also mark all elements that play a role in the expression of modality. The main components of our annotation scheme are:

- Trigger: the element conveying the modal value;
- Target: the expression in the scope of the trigger;
- Source of the event mention (speaker or writer);
- Source of the modality (agent or experiencer);

We decided to mark up two different sources to distinguish between the person who is producing the sentence with modal value and the person who is 'undergoing' the modality. For example, in 5, the source of the event mention is the speaker who states a certain fact in the sentence, while the source of the modality is the noun phrase *Os portugueses* who is the entity with the internal necessity triggered by the verb *necessitar* 'to need'. In these cases, the two sources do not refer to the same entity.

(5) Os portugueses *necessitam*, em média, de 180 contos por mês para a manutenção de uma família de quatro pessoas. 'Portuguese people need, on average, 180 thousand escudos per month to support a family of four people'.

The source of the event mention is usually not present in the sentence and in those cases it is not annotated. In Table 1 we detail what we define as source of the modality for each of the modal types. Although this information might seem obvious at first sight, this kind of explicitness has proved to be of great help for the annotators, since the source of the modality can be defined in different ways depending on the modal value and on the context. For the trigger we specify two attributes: 1) modal value; 2) polarity: an indication if the polarity of the modal value is positive or negative.
We consider polarity only in relation to the trigger and do not annotate the polarity of full sentences. We feel that negation is not an intrinsic part of modality but an external factor that often interacts with modality and that would deserve its own mark-up scheme (e.g. (Morante, 2010)). Or alternatively, create a scheme that captures both negation and modal triggers and combine them into one unified scheme as was done in the study of Baker et al. (2012 to appear). We didn't take such options for now, but considering that negation clearly changes the meaning of the modal expression in some cases, we do want to mark up this effect on the modality trigger in a separate feature. We kept the polarity feature as a simple binary feature where the positive value expresses the unmarked cases.
We have an additional field Comment in the annotation scheme to denote any difficulties, special cases and ambiguity. As modality is a complex phenomenon and not all cases are clear-cut, we addressed this problem, not by simplifying the annotation scheme or the guidelines, but to signal these cases explicitly in the comment field so that they can be studied in more detail in the future.
We annotate all elements that are part of the modal expression including verbs, nouns, adjectives, adverbs, prepositional phrases and clauses, but the annotation is limited to single sentences. For marking up the components in the scheme, we take a "min-max strategy" following Farkas et al. (2010). For the trigger we only annotate the minimum, the smallest possible unit (for example only the head noun in a noun phrase). For the target we annotate maximally and include all relevant parts. For the sources, we annotate full noun phrases or verbs (see section 3.1.). We consider the following categories as potential triggers when they denote modality of an event (and not of an entity): nouns, verbs, adverbs, adjectives that are part of a verbal phrase and the verb+prep combinations *ter de* 'must' and *haver de* 'have to'. Negation markers or auxiliary verbs are not considered part of a trigger. Example 6 shows which parts are annotated.

Modal value	Source of the modality
epistemic knowledge	who has the knowledge
epistemic belief	who has the belief
epistemic doubt	who has the doubt
epistemic possibility	who thinks something is possible
epistemic interrogative	who asks the question
part.-internal necessity	who has the necessity
part.-internal capacity	who has the capacity
deontic permission	who/what gives permission
deontic obligation	who/what makes the obligation
effort	who / what makes the effort
success	who or what was successful
volition	who wants to
evaluation	who evaluates something

Table 1: Type of source of modality for the different modal values.

(6) PCP *quer* esclarecimentos sobre Polis de Gondomar.
'The PCP wants clarification of the Gondomar Polis'
Trigger: quer
Target: esclarecimentos sobre Polis de Gondomar
Source of the modality: PCP
Modal value: volition
Polarity: positive

We also briefly mention what we do not consider as modal in our scheme. We do not annotate tense. So, we don't tag the past tense (although it provides certainty about the realization of an event), nor future (possibility) nor conditional (unless there is a conjunction introducing the conditional clause that we can consider a trigger). We also do not annotate declaratives, although they have an epistemic reading of belief (and even factuality) (Palmer, 1986). We consider, following Oliveira (1988), declaratives as representing the unmarked level of modality as they provide no evident trigger for the modal value. Evidentials are also not annotated as a separate value, and instead are marked as epistemic belief (supported by evidences). Finally, we do not annotate aspect. Verbs like *continuar a* 'continue', *passar a* 'change to', *acabar de* 'stop' signal the continuation of a state or event or a change of state, not a modality.

3.1. Annotation details

In the next part we discuss some complex cases and some language related problems like null subjects and clitics. As Portuguese is a null-subject language, the source of the modality is in some cases not expressed explicitly. In these cases we decided to tag as source of the modality the main verb that carries inflectional information pointing to the subject. In example 7 the verb is marked both as trigger and source, as *conseguimos* 'we managed'.

(7) *Conseguimos* voltar ao ponto zero!
'We managed to get back to square one!'
Trigger: Conseguimos
– Modal value: success
– Polarity: positive
Target: voltar ao ponto zero
Source of the modality: conseguimos

As target we mark nominal phrases, subordinate clauses and the verbal phrases in which all complements of the verb should be included. We include adverbial phrases only when they are structurally inside the scope of the target, because we aim to select only one continuous target if possible. In some cases the target is split in two parts like in example 8 where the prepositional phrase *No terreno das indústrias da cultura - cinema , livro , televisão* - is an essential part of the verbal phrase target[2].

(8) No terreno das indústrias da cultura - cinema, livro, televisão -, *arriscamo* -nos a ser dominados pelo mercado americano.
'On the terrain of cultural industry - film, books, tv -, we risk being dominated by the american market.'
Trigger: arriscamo
– Modal value: epistemic_possibility
– Polarity: positive
Target: No terreno das indústrias da cultura - cinema , livro , televisão - @a ser dominados pelo mercado americano
Source of the modality: -nos

This example also illustrates how we annotate verbs with clitic pronouns attached. The verb *arriscar-se* 'to risk' is inherently pronominal, meaning that it occurs with a clitic element (see *arriscamo-nos* 'we risk' in 8.). We do not consider clitics (reflexive or inherent to the verb form) as part of the trigger, but we do mark this element as source of the modality as it bears marks of person and number that refer to the omitted subject. We could consider to include the clitic as part of the trigger when the verb is intrinsically pronominal but experience shows that annotators often disagree about the inherent vs. reflexive interpretation of the context. So, we decide to always keep the clitic outside of the trigger and as the source of the modality here.

(9) "*Sabemos* sim , senhor Little Axe . . . "
'Yes we know, sir Little Axe ....'

[2]In this example we use the symbol @ to signal the discontinuity, in the corpus we use XML to mark this.

Figure 1: Screen shot of an sentence annotated with modal information in the MMAX2 environment. (Eng: Catarina never confessed her desire for her friend and this friend has always done everything not to know it.)

In rare cases it may happen that the modal event does not have a target expressed in the sentence. In example 9 the trigger *sabemos* 'we know' does not have a target as the sentence does not reveal what it is that is known. As the annotation is limited to single sentences, the target cannot be annotated in this case.
One substantial difficulty in the annotation is ambiguity as some cases express multiple modalities at the same time. We annotate this ambiguity by assigning multiple values in the field Modal value in the annotation scheme, as shown in 10.

(10) Médicos *exigem* pagamento de horas extras.
'Doctors require the payment of extra-hours'.
Trigger: exigem
– Modal value: deontic obligation; volition
Target: pagamento de horas extras
Source of the modality: Médicos

In case of questions and in case of exclamations in deontic contexts (imperatives), we denote the question mark and exclamation mark as being the trigger of the modality and the sentence itself as the target.

3.2. Annotation with MMAX2

For the final annotation of the corpus sample we used the MMAX2 annotation software tool (Müller and Strube, 2006). The MMAX2 software is platform-independent, written in java and can freely be downloaded from http://mmax2.sourceforge.net/. MMAX2 offers a visual interface to annotate sentences by marking up textual strings and creating links between the marked elements. The annotations are stored as stand-off XML. We implemented our modality annotation scheme in the MMAX2 environment. In the MMAX2 tool we consider modality as an event that has several marked elements ("markables") that participate in the modal event, namely the trigger, target, source of modality and source of event. We say that all these markables belong to the same modal event, which we call here a "set". The trigger markable has some specific features to be filled in: the modal value and its polarity. It also offers a text box for specifying ambiguity and additional comments. We chose MMAX2 as software as it offers a flexible definition of the elements to be marked. These markable are usually a textual string, but they can also consist of multiple discontinuous text parts. Markables can theoretically stretch over different sentences, they can overlap with each other and two different markables can cover exactly the same text string. We need this flexibility as we often have discontinuous targets or elements that express both a trigger and source of modality.

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE markables SYSTEM "markables.dtd">
<markable id="m_506" span="word_1200" mmax_level="modal" polarity="pos" modal_class="set_1" modal_value ="volition" type="trigger" / >
<markable id="m_505" span="word_1195" mmax_level="modal" modal_class="set_1" type="source_modality"/ >
<markable id="m_509" span="word_1201..word_1202" mmax_level="modal" modal_class="set_1" type="target" / >

<markable id="m_507" span="word_1206" mmax_level="modal" polarity="pos" modal_class="set_2" modal_value="effort" type="trigger" / >
<markable id="m_511" span="word_1209..word_1212" mmax_level="modal" modal_class="set_2" type="target"/ >
<<markable id="m_510" span="word_1204" mmax_level="modal" modal_class="set_2" type="source_modality" / >

<markable id="m_508" span="word_1212" mmax_level="modal" polarity="neg" modal_class="set_3" modal_value="epistemic_knowledge" type="trigger" / >
<markable id="m_512" span="word_1204" mmax_level="modal" modal_class="set_3" type="source_modality" / >
<markable id="m_513" span="word_1211" mmax_level="modal" modal_class="set_3" type="target" / >
</markables/ >
```

Figure 2: MMAX XML format for modality annotation for the sentence in figure 1.

We illustrate the annotation in MMAX2 with the sentence shown in figure 1. This sentence contains three modal triggers, the noun *desejo* 'desire' and the verbs *fez* 'did' and *saber* 'to know'. All parts of the sentence that are marked up as being involved in a modal event have a blue color in the MMAX2 visual interface. Figure 2 shows the XML markup that is produced by the MMAX2 tool as the result. The first trigger *desejo* (markable m_506 in figure 2) has modal value *volition* and a positive polarity. As MMMAX" produces stand-off XML, the words from the text are defined in a separate file and each word is numbered. In the modal annotation the feature *span* refers to the numbered words that are marked. The word *desejo* is denoted as *word_1200*. The target here is *pela amiga* 'for her friend'. The source of the modality is *Catarina*. Note that the sentence contains a negative element *nunca* 'never' which affects the declarative event expressed by the verb confessou 'confessed', and does not affect the positive polarity of the modal event *desejo*. The trigger, target and source are linked to the same modal event that is expressed by the feature *modal_class="set_1"*.

The second trigger *fez* has the modal value *effort* and is actually part of a semi-fixed multiword expression *fez de tudo* 'did everything' but we only mark the head node of the expression. The source of the modality is the pronoun *esta* 'this one_FEM', which is co-referent with the noun *amiga* of the previous clause. The target of *fez* is *para não o saber* 'not to know it'. The third trigger *saber* (m_508) has as modal value *epistemic knowledge* and has negative polarity caused by the word 'não'. The target of this modal event is *o seu desejo* that is expressed in the second clause with the referring pronoun *o*. This example illustrates how we deal with referring expressions. The annotation rule is to mark those elements that are present in the same clause as the trigger (so we annotate *o* and not *o seu desejo* in the previous clause).The pronoun *esta* is plays the role of source of the modality in two different modal events, and therefor is marked twice (m_510 and m_512).

The full annotation scheme is documented in the annotation guidelines (Hendrickx et al., 2012), where we detail all difficult cases, including how to deal with passive sentences, impersonals, and the limits of our annotation scheme.

4. Corpus Sample

We applied the annotation scheme for modality to a corpus sample of approximately 2000 sentences extracted from the written part of the Corpus de Referência do Português Contemporâneo (CRPC) [3] (Généreux et al., 2012), a highly diverse corpus of 312 million words covering a large variety of textual genres and Portuguese varieties. The creation of the CRPC started in 1988 by the Center of Linguistics of Lisbon University[4] and the aim was to create a diverse, balanced and up-to-date corpus of contemporary Portuguese. The written sub-part of the corpus consists of 310 million words, sampled from texts mostly after 1970 gathered from many different genres and domains such as scientific papers, technical reports, literary works, newspaper texts, parliament transcriptions and judicial documents. All texts have been automatically cleaned and linguistically annotated with POS-tags and lemmas. The CRPC also has a spoken part of 1.6 M words which we did not use in the current experiments. We extracted a corpus sample on the basis of a list of 40 Portuguese verbs that can express a modal meaning. We attempted to select equal sets of verbs that are associated to each type of modality. For example, the verbs *saber* 'know', *pensar* 'think', *crer* 'believe', *perceber* 'understand' and *julgar* 'judge' are generally associated with epistemic meaning and therefore chosen to trigger epistemic modality, while the verbs *permitir* 'allow', *obrigar* 'oblige', *exigir* 'require', *conceder* 'allow', *deixar* 'allow' are usually associated with deontic meaning. The modal verbs are used as a selection criterium to gather sentences containing at least one modal expression. However, the annotation of modality covers all modal elements present in the sentences, including nouns, adverbs and adjectives. We used the online interface for CRPC[5] to query for each verb lemma and retrieved the first 50 sentences from a randomly ordered list. We restricted the search query to cover only European Portuguese and excluded documents from Politics and Law to avoid formal language usage.

This method of selection of sentences for our corpus has some implications. First of all, querying for modal verbs

[3] CRPC:http://www.clul.ul.pt/sectores/linguistica_de_corpus/projecto_crpc.php

[4] CLUL: http://www.clul.ul.pt

[5] http://alfclul.clul.ul.pt/CQPweb

that are associated with certain modal values influences the frequencies of occurring modal values in the corpus sample. The frequencies found in the sample are by no means representative for the total CRPC corpus or for Portuguese in general. We expect to find a different distribution in another sub corpus or in non-sampled text. For example, in English biomedical text, modality is mostly expressed by adjectives, non-modal verbs, and adverbs (Thompson et al., 2008). Secondly, we chose to annotate single sentences and not full texts. Therefore we cannot treat co-referential relations between linguistic elements outside the sentence or recover the omitted element in an ellipsis, unless it is expressed within another clause in the sentence itself. Furthermore, annotating sentences without their original context leads to finding more ambiguities as a larger context could disambiguate their meaning. Even though our main findings are restricted to being only valid and representative for the corpus sample, we feel that the annotation of the corpus sample does give us an indication of what type of elements play a role and provide valuable insights in how modality is expressed in Portuguese.

5. Results

Here we present the general trends that we found in the annotated corpus sample. In a sample of 1946 sentences, we found 2377 triggers and 2509 modal values as we encountered 135 ambiguous cases. Table 2 presents the distribution of modal values in our sample: epistemic modality and deontic modality are the most frequent values, followed by volition. We notice that all different types of modal values occur a least more than 25 times. The triggers are mostly verbs as can be expected by our selection method. Some verbs that we had selected because we expected them to express modality, didn't live up up to our expectations. Of the 50 collected sentences of *falhar* 'to fail' only 9 occurrences had a modal meaning, for the verbs *garantir* 'to guarantee', *responsabilizar-se* 'to be responsible' and *falir* 'to fail' we only found 3, 2 and 1 examples respectively. On the other hand some modal verbs occurred much more often than the minimum of 50 and almost always expressed modality such as *poder* 'may/can' that occurred 210 times as a modal verb and 11 times without a modal meaning and *dever* 'must' which occurred 218 times with modal meaning and 10 times without. We also encountered several nominal triggers such as *tentativa* 'attempt' or *ambição* 'ambition' and adjectives that were part of a verbal phrase such as *dificil* 'difficult', *necessário* 'necessary', and *possível* 'possible'. We only encountered 5 different adverbial triggers that always co-occurred with another verbal trigger. These adverbs have the specific function of strengthening this verbal trigger. In example 11 the adverb *obrigatoriamente* 'obligatorily' is enforcing the deontic obligation that is expressed by the modal verb *ter de* 'must'.

(11) A feminilidade *tem de* ser *obrigatoriamente* excluída do feminismo? 'Does femininity have to obligatorily be excluded from feminism?'

The source of the event mention is usually the speaker or writer and only in 6% of the modal events, the source is

Modal value		Freq	%
Deontic		740	28,8
	obligation	581	22,7
	permission	159	6,2
Epistemic		739	28,8
	possibility	279	10,9
	knowledge	183	7,1
	belief	161	6,3
	interrogative	87	3,4
	doubt	29	1,1
Volition		396	15,4
Part internal		248	9,7
	capacity	126	4,9
	necessity	122	4,8
Evaluation		159	6,2
Success		119	4,6
Effort		110	4,3

Table 2: Frequency information about the modal values encountered in the corpus sample.

textually represented in the sentence. The source of the modality is present in 70% of the cases, in the other 30% it refers to the speaker or writer and so, the two sources refer to the same entity.

We encountered 450 modal events with a negative polarity and 84% of these were triggered by the word *não* 'no'. In most other cases the negation was conveyed by the modal verb itself like for example *vedar* 'to forbid' or *impedir* 'to prevent'. We consider all these cases as negative polarity.

The target constitute the elements in the sentence that are affected by the modality expressed in the trigger. Several types of elements can be targets: nominal phrases, verbal phrases, and subordinate clauses. In the majority of the cases the target is a subordinate clause or verbal phrase. However, in some cases, also main clauses can be targets. This occurs especially when there are two clauses, one of which has a parenthetical function as in example 12, where *como sabem* 'as you know' occurs within the main clause *só que no futebol tudo pode acontecer* 'in football everything can happen', being separated from it by two commas.

(12) Só que no futebol, como *sabem*, tudo pode acontecer, pelo que vamos esperar para ver. 'It is just that in football, as you know, everything can happen, so we will wait to see what happens'.

We conducted a small study to measure the inter-annotator agreement (IAA) for the task of modality annotation. Such a study gives some insight into the complexity of the task at hand, and the feasibility of the annotation scheme. Two linguists each annotated 50 sentences. We computed IAA using the kappa-statistic (Cohen, 1960) for each field in the annotation. For the Trigger the kappa value was .65 and for the accompanying Modal value a kappa of .85 was obtained, similar to reported IAA for English (Matsuyoshi et al., 2010).

6. Difficult cases

During annotation we encountered several difficult cases. It is exactly this type of cases that emphasize how important corpus-based research is for linguistic analysis as these provide us with insights in the complexity and the interactions that play a role in modality. Here we discuss several of these cases. Ambiguity of modal values is an obvious problem. In the corpus sample, the ambiguity between epistemic and deontic modality was very recurrent. In particular, we have found ambiguity between the sub values epistemic possibility and deontic permission. This ambiguity is mostly conveyed by the verb *permitir* 'to allow', which can be interpreted as expressing a possibility, or a permission, if there is some external factor allowing someone to do something or allowing something to happen.

(13) As condições climáticas *permitem* o desenvolvimento de árvores como abetos , pinheiros e outras plantas resinosas (coníferas). 'The climatic conditions permit the growth of trees such as spruce, pine and other coniferous plants (conifers)'.

In example 13, *permitir* 'to allow' is ambiguous: on the one hand, it expresses epistemic possibility, in the interpretation that the climate makes it possible for trees to grow. On the other hand, it expresses deontic permission if we interpret it in the meaning that the climate is a necessary condition for the growth of the trees. This ambiguity is the most frequent and occurs 34 times in the whole corpus. If compared to the frequency of epistemic possibility, the ambiguity has not a very high frequency value, but if compared to the frequency of deontic permission, we can see that it represents 21% of the total occurrences of deontic permission (Table 2). It is, then, important to keep the ambiguity annotated.

A second difficult case is polarity. We especially found some cases in which the identification of the polarity of the modal value was more complicated. In interrogative sentences, it is difficult to establish whether the polarity is positive or negative, because only the content of the interrogative can be negated, and not the question itself. However, we decided to avoid including a neutral value for polarity, and we treat the positive value as the default one, covering also these cases. Furthermore, when the trigger in the scope of the negative particle scopes over another modality trigger, it is possible that the negative polarity of the particle affects both triggers. For example, in 14 the negative adverb *nunca* 'never' applies to the trigger *conseguir* 'manage' with modal value success. The target of this verbal trigger includes another trigger *crer* 'to believe' with modal value epistemic belief. The general interpretation of the sentence is the negation of both modalities: success and belief, and so both modal values will be annotated with negative polarity of their modal values.

(14) É este um vício que sempre atinge os míseros: *nunca conseguir crer* na felicidade! 'And this is a vice that always affects poor people: to never manage to believe in happiness!'

Some cases are more difficult to capture in our annotation. For example, if the negative particle is contained in the target, we have no means to annotate it, since we only describe negation when related to a trigger with modal value. Contexts with two negative particles, both in the trigger and the target, raise the same issue. In 15, the epistemic modality has negative polarity expressed by the adjective itself *impossível* 'impossible' and the target has a negative adverb scoping over the participant-internal modality (*capacidade* 'capacity'). The overall interpretation of the sentence is that the entity has in fact effective internal capacity, but our annotation processes each trigger independently and does not capture the overall positive polarity of the sentence.

(15) Era *impossível* dizer que não tínhamos *capacidade* para crer, para amar ou para adorar. 'It was impossible to say that we had no capacity to believe, to love or to worship'.

A recurrent situation during the annotation was that there were two triggers for two different modal values and the second was also part of the target of the first trigger, influencing the meaning of the second trigger. In the corpus sample there were several examples in which the value epistemic possibility influenced the certainty of other values, specifically of the values evaluation and deontic obligation. Other values that influence the certainty of the modal verb in the sentence are the values epistemic interrogative and epistemic doubt. We illustrate this influence in example 16 where the trigger *pode* 'can' influenced the certainty of the modal value evaluation of the second trigger *difícil* 'difficult'.

(16) Se o aluno se perde, *pode* ser *difícil* voltar a apanhar. 'If the student looses himself, it can be difficult for him to catch up again'.

In the current annotation scheme we denote ambiguity of the modal values. However, in some cases we encounter a structural ambiguity that not only influences the modal value, but also the components involve change and this type of ambiguity can not be captured in our current annotation scheme. In the example 17 two interpretations and therefore two annotations are possible. In the first reading where *ter de* expresses deontic obligation, the source of the modality is the speaker/writer and the subject *A* is part of the target. In the second reading, the sentence expresses a participant internal necessity (internal necessity to A) and the source of the modality is therefore *A* . Without a larger context the ambiguity of this sentence cannot be resolved. We use the field Comment to mark up these cases and name them "structural ambiguity". We counted the number of times this type of ambiguity occurred with the verb *ter de/que* 'must' in our corpus and we found 70 examples. Currently the scheme has no way of annotating these structural ambiguities. However, in further research we plan to look at these cases and see whether we can find a solution for the annotation.

(17) A tem de ser feito. 'It must be done'

Ambiguity	option 1	option 2
Trigger	tem de	tem de
Modal value	d. obligation	part. int. necessity
Target	A@ser feito	ser feito
Source	*speaker*	A

7. Conclusion

In this article we presented a scheme for the annotation of modality in Portuguese and we discussed the details of the scheme, difficult cases and limitations. We presented a corpus sample annotated with modality and our main findings in the data. As we have shown there are several complex cases that deserve a more in-depth study. We also would like to take a closer look at some of the modal values, in particular the modal value *evaluation*: it has proved difficult to establish exactly which triggers carry this modal value and in which contexts, so there is the need to refine the scope of this value. Furthermore, the interaction of evaluation and polarity seems to be more complex. While with other values, negative polarity negates the value itself, with evaluation the presence of a negative element does not negate that there is an evaluation, but rather changes the kind of evaluation. The scheme for the annotation of modality in corpora has so far only been applied to sentences, but in future work we would like to apply it to full texts. We also plan to study the applicability of the annotation scheme to spoken material as in oral language, modality can also be expressed by extra-linguistic elements or by typical oral lexical expressions.

Acknowledgement

We would like to thank Agostinho Salgueiro for his annotation work. This work is financed by the Fundação para a Ciência e a Tecnologia (FCT) for the project PEst-OE/LIN/UI0214/2012, by the project METANET4U (CIP-ICT-PSP 270893) and by the FCT Doctoral program Ciência 2007/2008.

8. References

Kathrin Baker, Michael Bloodgood, Bonnie Dorr, Nathaniel W. Filardo, Lori Levin, and Christine Piatko. 2010. A modality lexicon and its use in automatic tagging. In *Proceedings of LREC'10*, Valletta, Malta. ELRA.

Kathryn Baker, Bonnie Dorr, Michael Bloodgood, Chris Callison-Burch, Nathaniel Filardo, Christine Piatko, Lori Levin, and Scott Miller. 2012, to appear. Use of Modality and Negation in Semantically-Informed Syntactic MT. *Computational Linguistics*.

J. L. Bybee, R. Perkins, and W. Pagliuca. 1994. *The evolution of grammar: Tense, aspect and modality in the languages of the world*. University of Chicago Press, Chicago.

J. Cohen. 1960. A coefficient of agreement for nominal scales. *Education and Psychological Measurement*, 20:37–46.

Richárd Farkas, Veronika Vincze, György Móra, János Csirik, and György Szarvas. 2010. The conll-2010 shared task: Learning to detect hedges and their scope in natural language text. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, pages 1–12. ACL.

Michel Généreux, Iris Hendrickx, and Amália Mendes, 2012. *Proceedings of PROPOR 2012, LNAI*, volume 7243, chapter Introducing the Reference Corpus of Contemporary Portuguese. Springer.

Iris Hendrickx, Amalia Mendes, Silvia Mencarelli, and Agostinho Salgueiro, 2012. *Modality Annotation Manual, version 1.0*. Centro de Linguística da Universidade de Lisboa.

Suguru Matsuyoshi, Megumi Eguchi, Chitose Sao, Koji Murakami, Kentaro Inui, and Yuji Matsumoto. 2010. Annotating event mentions in text with modality, focus, and source information. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. ELRA.

Marjorie McShane, Sergei Nirenburg, Stephen Beale, and Thomas O'Hara. 2005. Semantically rich human-aided machine annotation. In *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*, pages 68–75. ACL.

R. Morante. 2010. Descriptive analysis of negation cues in biomedical texts. In *Proceedings of the Seventh International Language Resources and Evaluation (LREC'10)*, pages 1429–1436, Valletta, Malta.

Christoph Müller and Michael Strube. 2006. Multi-level annotation of linguistic data with MMAX2. In *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*, pages 197–214. Peter Lang.

Sergei Nirenburg and Marge McShane. 2008. Annotating modality. Technical report, University of Maryland, Baltimore County, March 19, 2008.

F. Oliveira. 1988. *Para uma semántica e pragmática de DEVER e PODER*. Ph.D. thesis, Universidade do Porto.

F. R. Palmer. 1986. *Mood and Modality*. Cambridge textbooks in linguistics. Cambridge University Press.

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

György Szarvaz, Veronika Vincze, Ricárd Farkas, and János Csirik, 2008. *The BioScope corpus: annotation for negation, uncertainty and their scope in biomedical texts*, pages 38–45. ACL.

P. Thompson, G. Venturi, J. Mcnaught, S. Montemagni, and S. Ananiadou. 2008. Categorising modality in biomedical texts. In *Proceedings of the LREC 2008 Workshop on Building and Evaluating Resources for Biomedical Text Mining 2008*.

J. Van der Auwera and V. Plungian. 1998. Modality's semantic map. *Linguistic Typology*, pages 79–124.

Veronika Vincze, György Szarvas, Móra György, Tomoko Ohta, and Richárd Farkas. 2011. Linguistic scope-based and biological event-based speculation and negation annotations in the bioscope and genia event corpora. *Journal of Biomedical Semantics*, 2(5).

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39:165–210.