

Modality annotation for Portuguese: from manual annotation to automatic labeling

AMÁLIA MENDES¹, IRIS HENDRICKX^{1,2}, LUCIANA ÁVILA^{1,5}, PAULO QUARESMA^{3,4}, TERESA GONÇALVES³ AND JOÃO SEQUEIRA³,

¹*Center for Linguistics, University of Lisbon, Portugal*

²*Center for Language Studies/Center for Language and Speech Technology, Radboud University, Nijmegen, The Netherlands*

³*Department of Informatics, University of Évora, Portugal*

⁴*L2F – Spoken Language Systems Laboratory, INESC-ID, Portugal*

⁵*Universidade Federal de Viçosa, Brazil*

Abstract

We investigate modality in Portuguese and we combine a linguistic perspective with an application-oriented perspective on modality. We design an annotation scheme reflecting theoretical linguistic concepts and apply this schema to a small corpus sample to show how the scheme deals with real world language usage. We present two schemas for Portuguese, one for spoken Brazilian Portuguese and one for written European Portuguese. Furthermore, we use the annotated data not only to study the linguistic phenomena of modality, but also to train a practical text mining tool to detect modality in text automatically. The modality tagger uses a machine learning classifier trained on automatically extracted features from a syntactic parser. As we only have a small annotated sample available, the tagger was evaluated on 11 modal verbs that are frequent in our corpus and that denote more than one modal meaning. Finally, we discuss several valuable insights into the complexity of the semantic concept of modality that derive from the process of manual annotation of the corpus and from the analysis of the results of the automatic labeling: ambiguity and the semantic and syntactic

properties typically associated to one modal meaning in context, and also the interaction of modality with negation and focus. The knowledge gained from the manual annotation task leads us to propose a new unified scheme for modality that applies to the two Portuguese varieties and covers both written and spoken data.

1 Introduction

There has been a growing interest in text mining applications that can automatically detect opinions, facts and sentiments in texts. Many of the current opinion or sentiment mining applications use a crude division between negative, neutral and positive sentiments. Modality, defined from a linguistic perspective as the speaker's attitude towards the proposition in the text (Palmer, 1986) offers a theoretical framework to make more fine-grained distinctions between different attitudes. For example, one can detect whether the speaker expresses his or someone else's commitment, hope, belief or knowledge about a certain proposition.

In this paper we aim to combine a linguistic perspective with a practical and application-oriented perspective on modality. On the one hand we design an annotation scheme reflecting theoretical linguistic concepts and apply this schema to corpus data to fit with real world language usage. On the other hand we use the annotated data not only to study the linguistic phenomena of modality, but to train a practical text mining tool to detect modality in text automatically.

The concept of modality is related to factivity, evidentiality, hedging, and is consequently grounded on notions of subjectivity and attitude. The study of modality relates to the question of how subjective language use is. How many of our utterances and sentences are modalized? How is subjectivity and attitude grammaticalized in the sentence through modal markers? Are certain genres more prone to subjectivity? And how successful can we be in dealing with modality in automatic processing systems?

These questions depart from a purely formal approach to language and move into a cognitively grounded discourse perspective. While certain linguistic aspects such as morphology, lexicography and syntax have been addressed in interdisciplinary studies combining theoretical and computational approaches that have produced a body of available resources, other aspects such as semantics and discourse are still lacking the same level of attention. Although modality has been intensively explored from a theoretical linguistic perspective, there are few theoretical studies relying on corpus data and no concerted effort to create

a resource covering this topic.

In this paper we aim to contribute to filling this gap and provide an annotation schema for modality based on linguistic theory but firmly grounded by applying and revising the schema based on corpus annotation. This schema is also evaluated from a computational perspective as we use the annotated data to train an automatic modality tagger intended to be part of a Text Mining application. Moreover, many of the existing efforts in this domain are applied to written texts, while we discuss here the annotation of both written and spoken data. Furthermore, while most work on human and automatic annotation of modality is focused on English, we provide feedback on another global language, Portuguese, and address language diversity by covering two national varieties of Portuguese (Brazilian and European).

Corpus annotation does not solely provide the basis for text mining applications, but sheds new light on linguistic issues. Indeed, we believe that corpus annotation triggers research questions left untouched when exclusively assessing constructed samples and allow to redefine categories after assessing the contexts. First, the specific task of labeling each modal marker in context challenges the usual theoretical perspectives that center on specific lexical items, mainly semi-auxiliary verbs, and leads us to reflect on the diversity of modal markers, providing the necessary data for an approach at the crossroad with typological linguistics. Second, this has implications over the modal types traditionally described in linguistics. It is necessary to address the complexity of corpus data and keep a balance between achieving accuracy and preventing the proliferation of labels. This inevitably requires an investigation of the role played by modal marker's ambiguity. Third, the analysis of corpus data will also provide empirical insights into theoretical issues to modality: modal logic (Kratzer, 2013) and discourse-focused modality (Bybee and Fleischman, 2013); propositional modality and event modality (Palmer, 2001). Finally, embedding the annotation in actual corpus data reveals the interaction of modal markers with systems such as negation and focus, which falls outside modality but may affect the modal interpretation.

Computationally, the automatic identification and interpretation of modalized statements is a prime concern in a large number of applications, especially with the recent attention to opinion mining and social networks. As the vast amount of digitally available data keeps growing, so does the demand to automatically extract relevant information. 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, and to distinguish between factual and

probable information (Saurí et al., 2006), to detect uncertainty, speculation and negation, especially in biomedical text mining (Baker et al., 2010, Matsuyoshi et al., 2010, Szarvas et al., 2008). Modality detection is therefore also clearly linked to the current trend in NLP on sentiment analysis and opinion mining.

We will report on the two modality annotation schemes designed for Portuguese, each addressing a specific national variety and a specific text type. The European Portuguese (EP) scheme was envisaged as genre independent and tested on written data, while the Brazilian Portuguese (BP) scheme was designed for spontaneous speech, and modality is taken as the evaluation or the point of view of a conceptualizer towards the locutory material in a given utterance in a communicative act. Both schemes cover different grammatical categories that convey modal values, such as adjectives, adverbs, nouns and verbs which we will denote as *modal trigger* in the rest of this paper. The EP scheme has been manually applied to a small written sample of 160K tokens (Hendrickx et al., 2012a) while the BP scheme was manually applied to a spoken sample of the CORAL-BRASIL spoken corpus (Ávila, 2014).

The EP annotated sample was used as the basis for the development of a software tool that can automatically detect modal triggers and their scope in text. Such automatic tool has many practical applications in the area of automatic document understanding or information extraction. The first experiment has been to automatically label a subset of 11 highly frequent and ambiguous modal verbs. For instance, the Portuguese verb *poder* is highly ambiguous and can express multiple modal meanings: “to be possible”, “to be able” or “to give permission”. This polysemy increases the level of difficulty of the automatic annotation task. To create the modality tagger, we first automatically assign lemmas, POS and syntactic tags, we then automatically identify modal triggers and apply a machine learning approach to attribute a modal value to the triggers, comparing the results with our gold (manually annotated) labeling.

Constructing an automatic modality tagger requires a data set with labeled examples to train and evaluate the tagger. As we are currently in need of a suitable data set, one of the goals of the current experiments is to develop an automatic modality tagger on a small manually labeled sample that can later be applied (semi-automatically) to generate a larger data set.

Our paper first discusses related work on modality, its annotation and automatic labeling in section 2. The two schemes are presented in section 3. The automatic labeling tool trained over the EP corpus is presented in section 4. In section 5 we discuss several valuable insights

into the complexity of the semantic concept of modality that the process of manual annotation of the corpus, its linguistic exploration and the analysis of the results of the automatic labeling have provided us. In the last two sections we discuss the next steps that we aim to take such as using a unified approach to modality annotation presented in section 6 and we conclude in section 7.

2 Related work

Modality is traditionally grounded in epistemic modality, which is concerned with the truth-value of the proposition: whether the proposition is considered by the speaker (or another entity) as certain, uncertain, possible, probable etc. In languages with no modality-exclusive morphological system, such as English or Portuguese, research has mainly focused on semi-auxiliary modal verbs, e.g., can, may, must. For instance one can have different perspectives on the proposition (Ana buy a house), such as: Ana may buy a house, Ana must buy a house, Ana can buy a house. Other modal types besides epistemic modality have been proposed that reflect some degree of subjectivity towards the proposition (or towards the event). Although terminologies vary considerably, the concepts at hand are relatively constant: they include notions of obligation and permission, of capacity and necessity, of evidentiality (what is the source of information), wish and evaluation. This broader scope of modal types is reflected in the growing attention to a large set of linguistic categories that may convey modal values, such as tense affixes, mood, auxiliary verbs, main verbs, adjectives, adverbs and multi-word expressions. In fact, several lexical items conveying modality may occur in the same context, and, furthermore, many of these items are ambiguous between one or more modal values. The complexity of the phenomenon does not undermine its importance in language to convey attitudinal information, such as belief, uncertainty, factuality and evidentiality.

Most of the research on modality has been based on constructed examples, although the importance of looking at modal items in context is increasingly acknowledged, on a par with a clear trend in Corpus Linguistics and Natural Language Processing (NLP) to go beyond the part-of-speech (POS) and syntactic levels and to include semantics, pragmatics and supra-sentential information.

This interest gave rise to some proposals for the annotation of modality in corpora, mainly for the English language. The annotation schemes covering modality differ greatly in their objectives and in the nature of the concepts that are labeled (Nissim et al., 2013). According to

Nissim and Pietandrea (2015),

computationally, the automatic identification and interpretation of modalized statements is a prime concern in a large number of applications, especially with the recent attention to opinion mining and social networks (pp. vii)

but

the computational linguistics community is still far from having developed working, shared standards for converting modality-related issues into annotation categories. (pp. vii)

Modality may be one aspect of the semantic information encoded in the properties of events (cf. (Baker et al., 2010, Matsuyoshi et al., 2010, Nirenburg and McShane, 2008, Song et al., 2015, Szarvas et al., 2008)) or it can be the core of the scheme. For instance Rubinstein et al. (2013) use a restricted notion of modality and establish conditions for an expression to be considered modal, such as the requirement for a propositional argument. In other cases, modal values are included in annotation schemes that cover both factuality and modality, as in the work of Saurí et al. (2006) that distinguish between factual and probable information, while Lee et al. (2015) specifically address factuality, rated on a scale of -3 (certainly did not happen) to 3 (certainly did).

Different annotation schemes also diverge in what textual elements they annotate: the modal value may be attributed globally to the sentence/event or it can be encoded on specific lexical items. The applied nature of these studies leads to detailed description of the textual elements involved in the expression of modality and the roles they have: most schemes identify a modal trigger, the subject of the modality (source) and the elements in the scope of the modal trigger (target/scope/focus).

The availability of large sets of data annotated with modality provides important insights on the interaction between modality and other linguistic systems, such as negation (Morante and Sporleder, 2012) and focus (Mendes et al., 2013, Moreira, 2005). Another application of these data sets are experiments in the automatic annotation of modal values for information extraction and data mining. In BioNLP, Miwa et al. (2012) annotated pre-recognized events with the epistemic value “level of certainty” and attain F-measures of 74,9 for “low confidence” and 66,5 for “high but not complete confidence”. The factuality-oriented scheme presented by Saurí et al. (2006) has been applied in an experiment of automatic identification of events and their modal features in text, and attains 97.04 accuracy with the EviTA tool. A specific task for detecting uncertainty through the use of hedging clues was

organized at CoNLL2010 (Farkas et al., 2010). Contrary to these approaches, Baker et al. (2010) include both trigger and target in their experiments, and attain 86% precision with a structure-based tagger over 249 modality-tagged sentences from the English side of the NIST 09 MTEval training sentences. The approach of Diab et al. (2009) covers modality but is essentially geared towards the identification of belief. Contrary to our own approach, the authors do not take into consideration the polysemy of the auxiliary verbs and only encode the epistemic value, although they do report that the verbs may be deontic in some contexts. This experiment has been extended to other modality values (ability, effort, intention, success and want) (Prabhakaran et al., 2012). The work of Ruppenhofer and Rehbein (2012) focuses on English modal verbs. The five English modal verbs (*can/could, may/might, must, ought, shall/should*) were first identified in texts and their modal values were predicted by training a maximum entropy classifier on features extracted from the training set. The classifier achieved an improvement of the most frequent sense baseline for all verbs but *must*, and accuracy numbers between 68.7 and 93.5. Our experiments presented in Section 4 are closely related to this type of automatic modal verb labeling. This brief overview of related work has shown the diversity of objectives, annotation choices and evaluation metrics, which makes it difficult to compare results.

3 Annotation schemes

We present two separate modality annotation schemes for Portuguese. Although they apply to different varieties of Portuguese, the general structure of the schemes is very similar. For instance, the range of phenomena that are not considered as modality, and are not annotated, is almost equivalent in both schemes:

- Although tense contributes to modality and interacts with the modal values in the sentences, it is not annotated. The schemes do not tag the past tense (although it provides certainty about the realization of an event), nor future (possibility) nor conditional tense¹.
- Declarative clauses with no modal triggers convey an assertion and may have an epistemic reading of belief (and even factuality) (Palmer, 1986), but both schemes consider the declarative sentence type as non-modal or as representing the unmarked level of modality (John Lyons, 1977, Oliveira, 1988).

¹Note however that the conjunction introducing the conditional clause is considered a lexical trigger.

- The Evidential modal type, i.e., contexts where the source provides some kind of evidence for his belief (for instance: *I believe it is a good movie, based on the reviews*), are also not annotated as a separate value, and instead are marked as epistemic belief.
- Aspectual verbs as *continuar a* ‘continue to’, *passar a* ‘change to’, *acabar de* ‘stop’ signal the continuation of a state or event or a change of state, and are not annotated.
- We only annotate events, not entities. This proved especially important with the modal value evaluation, since many cases of evaluation have scope over entities.

3.1 Annotation scheme for written European Portuguese

The annotation scheme components comprise the trigger (which is the lexical clue conveying the modal value), its target, the source of the event mention (speaker or writer) and the source of the modality. The source of the event mention and the source of the modality are in many cases the same entity: the speaker/writer produces a discourse/text unit where he/she states his/her belief or doubt or the possibility that something may happen. However, they may also be different entities when the text presents the views of someone else than its producer. The annotation was performed over each trigger, and not over the global interpretation of the sentence. The trigger receives an attribute *modal value*, while both trigger and target are marked for polarity. This feature describes the positive or negative polarity of the trigger and not the polarity of full sentences. A trigger may be marked with negative polarity by a negation adverb, or the negative polarity can be expressed morphologically by an affix (*improbable*) or by the lexical verbal form itself (*proibir* ‘forbid’).

The choice of the size of the components to be annotated is a challenge. To achieve some level of consistency, we were inspired by the “min-max strategy” presented by Farkas et al. (2010). The trigger is annotated as the smallest possible unit (for example only the head noun in a noun phrase), while the target is annotated maximally and include all relevant parts. Adverbial adjuncts are included only when they are structurally inside the scope of the target, in order to avoid a proliferation of discontinuous targets. It is nevertheless difficult to establish exactly what is to be considered relevant and annotators still differ in this regard. For the sources, we annotate full noun phrases or verbs. In fact, as Portuguese is a null-subject language, the source of the modality is frequently not expressed explicitly in the sentence. In these cases we decided to tag the main verb that carries inflectional information pointing to the subject, or the clitic in cases of intrinsically pronomi-

nal verbs such as *arriscamo-nos* (verb-clitic) ‘we risk’ as source of the modality. We refer to annotation guidelines (Hendrickx et al., 2012b) for more details.

Modal verbs may have more than one meaning and it is sometimes difficult to distinguish between those modal values, even when the annotator takes into consideration a larger context. To address this issue, the scheme includes an Ambiguity field, where the annotators can write down secondary meanings when present in a specific context. The annotator chooses the most salient meaning as the main modal value. In general such ambiguity is caused by the intrinsic ambiguity of the modal trigger that has multiple meanings. In some rare cases, both options are equally salient, and the annotator has to choose randomly. We discuss ambiguity in more detail in section 5. The annotation scheme was further enriched with Focus information to address the interaction between exclusive adverbs and modal triggers (Mendes et al., 2013). The scheme is similar to the representation of modality (Nirenburg and McShane, 2008) used in OntoSem (McShane et al., 2005), a text processing environment that produces formal text meaning representations.

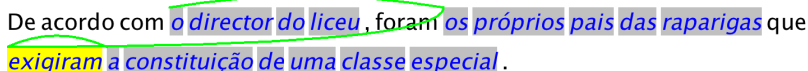
Our scheme covers a total of 13 modal values: epistemic and its sub-values (knowledge, belief, doubt, possibility and interrogative), deontic and sub-values (obligation, permission), participant-internal and sub-values (capacity, necessity), volition, evaluation, effort and success. The typology follows closely linguistic proposals such as Palmer (1986). The inclusion of participant-internal values follows directly from the typology of van der Auwera and Plungian (1998) (these values match what is called dynamic modality in other typologies (Palmer, 1986)). However, contrary to van der Auwera and Plungian (1998), the scheme doesn’t consider participant-external modality as an independent type, but rather as a sub type of deontic modality². Other values, such as effort and success, are inspired in other annotation schemes for modality (Baker et al., 2010).

We applied this scheme to a corpus of 160K extracted from the EP written sub-part of the Reference Corpus of Contemporary Portuguese (CRPC) (G en ereux et al., 2012), excluding however documents from Politics and Law to avoid formal language usage. A list of 50 pre-selected modal verbs were used as selection criteria to randomly gather 50 sentences per modal verb. However, the annotation of modality covers all modal elements present in the sentences, including nouns, adverbs and adjectives. One of the implications of the corpus selection method is that the frequencies found in the sample are by no means

²See Section 6 for an alternative approach.

representative for the total CRPC corpus or for Portuguese in general. Nevertheless, we feel that the results provide an indication of what type of elements play a role and provide valuable insights in how modality is expressed in Portuguese.

The editor used for this task is the MMAX2 annotation software tool (Müller and Strube, 2006) which is platform-independent, written in java and freely downloadable from <http://mmax2.sourceforge.net/>. MMAX2 offers a visual interface to annotate sentences by marking textual strings and creating links between the marked elements. MMAX2 allows to mark non-contiguous elements and produces stand-off XML annotation. The elements of the annotation consist of markables (namely the trigger, target, source of modality and source of event) that are linked to the same modal event. We present a screenshot of modal annotation in the MMAX2 editor in Figure 1. The trigger is marked in yellow, and each component of the set is linked to the trigger (source of the event, source of the modality, target).



De acordo com o director do liceu, foram os próprios pais das raparigas que exigiram a constituição de uma classe especial.

FIGURE 1 Screen-shot of MMAX2 annotation tool. (Eng: According to the high-school director, it were the girls' parents themselves who demanded the creation of a special class.)

The annotation was done by one annotator and all difficult cases were discussed with a second annotator and included in the guidelines. The scheme underwent several adaptations during the annotation process, to answer specific cases, to achieve consistency and also to revise the list of modal values according to the data. For example, the modal value commissive (Palmer, 1986) had few occurrences and its contexts were later included in the modal value deontic obligation.

A study to measure the inter-annotator agreement (IAA) for the task of modality annotation was conducted over a small subset of our data by targeting two specific features of our scheme: trigger identification and modal value attribution. Our goal was to obtain some insight into the complexity of the task 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 the reported IAA for English (Matsuyoshi et al., 2010).

We annotated 3183 lexical triggers in the 160K corpus. We present the list of values and sub-values in table 1, together with their frequency in our golden set. The annotation has been revised in several occasions and the numbers are updated in comparison to those presented in Hendrickx et al. (2012a) and the manual annotation guidelines reflect these updates (Hendrickx et al., 2012b).

Main modal values	Sub-values	Freq	%
Epistemic	knowledge	198	5,9
	belief	310	9,2
	doubt	15	0,5
	possibility	507	15,1
	interrogative	68	2
Deontic	obligation	543	16,2
	permission	221	6,6
Participant-internal	capacity	214	6,4
	necessity	129	3,8
Evaluation		75	2,2
Volition		415	12,3
Effort		358	10,7
Success		299	8,9
Total		3352	100

TABLE 1 Modal values and frequencies in the EP corpus

Epistemic modality and deontic modality, the core values of most linguistic typologies, are indeed the most frequent values, followed by volition. The triggers are mostly verbs as can be expected by our selection method. We also encountered several nominal triggers such as *tentativa* ‘attempt’ or *ambição* ‘ambition’ and adjectives that were part of a verbal phrase such as *difícil* ‘difficult’, *necessário* ‘necessary’, and *possível* ‘possible’. We only encountered 5 different adverbial triggers that always co-occur with another verbal trigger and have the specific function of strengthening this verbal trigger. The source of the event mention is lexically marked in only 6% of the modal events and it usually corresponds to the speaker or writer. When the source of the modality is not lexically marked, it refers to the speaker or writer and so, the two sources refer to the same entity.

There are 3183 triggers in our annotated corpus, but a total of 3352 modal values attributed, due to ambiguous cases. Sometimes the anno-

tators decided that multiple readings of the same sentence were plausible, and this was separately denoted.

We provide in (1) and (2) examples of modal uses of the verb *dever* ‘must/ may’. In (1), *dever* has an epistemic reading, stating that the proposition is probable (the adverb *provavelmente* ‘probably’ is also a modal trigger), while in (2), it has a deontic obligation reading. In (2) we also provide the total description of the components of the sentence annotation. There is no overt source of the modality (in these cases a verb form with inflection marks of the subject is tagged as the source). Cases marked as ambiguous are illustrated in (3): the context is ambiguous between a deontic obligation reading and an epistemic possibility reading (the new phases of the metro network should be vs. will probably be). The annotator selects what he considers to be the primary value and marks the ambiguity in the Ambiguity component.

- (1) E que *deverá* provavelmente ganhar, a julgar pelo acolhimento que o seu programa recebe.
 ‘And that shall probably win, judging by the reception that his program receives.’
- (2) Procurei informação sobre o que *deveria* fazer para reservar uma mesa para outro espectáculo.
 ‘[I] looked for information about what [I] should do to make a reservation to another show.’
 Trigger: *deveria*
 Modal value: deontic_obligation
 Polarity: positive
 Target: fazer para reservar uma mesa para outro espectáculo
 Source of the modality: procurei
 Source of the event: writer
 Ambiguity: none

- (3) É que, depois de tantos "passos de gigante", criou-se a natural expectativa de ver quais serão as novas fases da rede do metro lisboeta - que julgamos *deverem* ser, por um lado, de consolidação (...) e, por outro, de expansão (...).
 'It's just that, after so many "giants steps", there is a natural expectation about which will be the new phases of the metro network in Lisbon - which we believe should be / will probably be, on the one hand, consolidation and, in the other hand, expansion.'
 Trigger: *deverem*
 Modal value: Epistemic_possibility
 Polarity: positive
 Target: ser, por um lado, de consolidação (...) e, por outro, de expansão (...)
 Source of the modality: julgamos
 Source of the event: writer
 Ambiguity: Deontic_obligation

3.2 Annotation scheme for spoken Brazilian Portuguese

While the modal scheme for EP has been designed and applied to written texts, the modal scheme for BP is designed for spontaneous speech, which according to Cresti and Scarano (1998), is governed by an illocutionary principle not found in written texts, as well as specific informational articulations. It is more theory oriented: modality is understood in enunciative terms (Bally, 1932), that is, it is a semantic category applied to a conceptualizer's linguistic production, which qualifies and relativises the uttered locutive material in terms of degree of certainty, possibility, necessity, capability, and volition. The scheme follows the Language Into Act Theory (Cresti, 2000), which takes the utterance as its reference unit, and considers the scope of the modality to be the information unit (Tucci, 2008), whose locutory material does not necessarily express a proposition.

The scheme was applied to the Brazilian Portuguese spontaneous speech corpus C-ORAL-BRASIL I (Raso and Mello, 2012). This corpus follows the same architecture as the European Romance spontaneous speech 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 I comprises 200 texts of approximately 1,500 words each, proportionally distributed into dialogues, conversations and monologues. The corpus follows the CHILDES-CLAN³ transcription format to which prosodic

³CHILDES - Child Language Data Exchange System, at:

annotation is added, marking tone unit and utterance boundaries. The entire corpus is speech to text aligned with the WinPitch software (Martin, 2004).

In this particular study an annotated sample from the C-ORAL-BRASIL I was analyzed. It covers 20 texts, totaling 31,318 words, 5,484 utterances and 9,825 tone units. Firstly, the identification and classification of modal markers was undertaken by three annotators working independently; the codification was then qualitatively validated through group discussions involving the research group coordinator and her students. The search for modal markers was performed manually, through qualitative transcription examination, supported by the WinPitch text-to-audio aligned files and their concomitant examination through the software interface that allows speech signal listening as well as transcription and prosodic parameter visualization. The data were organized in a table containing the modal markers, the tone unit in which they occur, the type of information unit they are inserted in, the file they belong to, and any qualitative information deemed relevant.

The BP scheme (based on the latest revision of the guidelines in Ávila (2014)) uses a three-category scheme of epistemic, deontic and dynamic modality, inspired by Palmer (1986). Epistemic modality carries seven sub-values: knowledge, belief, possibility, probability, necessity (the conceptualizer presents what is said as a necessity, based on previous knowledge (*só pode ser doido* ‘he can only be crazy, he has to be crazy’)) and verification (the conceptualizer regards a state of affair as uncertain (*olha aí se nu tem ninguém* ‘check over there if there is no one’)). Deontic modality encompasses four sub-values: obligation, permission, prohibition and necessity (the conceptualizer expresses his or someone else’s needs). Finally, dynamic modality comprises the sub-values ability and volition/intention. In the sample, we found 1,088 modal markers (lexical and grammatical, excluded the conditional constructions) and from these we tagged 781 lexical triggers. The distribution of modal values and sub-values in the sub-corpus is presented in table 2.

Both schemes agree in not marking mood and tense. Nor do these schemes address factuality or a larger category of subjectivity and emotion. Due to their work on speech, Ávila and Melo (2013) and Ávila (2014) also distinguish modality, which is marked lexically and grammatically, from the pragmatic categories of illocution and attitude, which are carried by prosodic cues. As the three categories are often confused in their definition in the linguistic studies tradition, Mello

Main modal values	Sub-values	Freq	%
Epistemic	knowledge	100	14,9
	belief	228	40
	possibility	15	2,2
	probability	24	3,6
	necessity	15	2,2
	verification	14	2
Deontic	obligation	96	14,3
	permission	70	10,4
	prohibition	6	0,9
	necessity	17	2,5
Dynamic	ability	17	2,5
	volition	69	10,3
Total		671	100

TABLE 2 Modal values and frequencies in the BP spoken corpus

and Raso (2011), through experimental investigation and observation of empirical data, suggest that modality is restricted to the semantic domain, although interrelated and projected into the pragmatic one. The same illocution can be modalized in different ways and performed with different attitudes, without affecting the illocutionary level.

The EP and BP schemes share the same components: the Trigger is the lexical item that carries modality; the Source of the modality is the conceptualizer, i.e., the individual whose perspective and view point is being reported (this might be the speaker, the addressee, or another entity in the discourse); Source of the event mention is the producer of the text or the speaker; the Target is the expression in the scope of the trigger. The BP scheme also considers a Target-dependent component to encompass the cases in which the target, in a given utterance, is not explicit, but it can be recoverable in the referential chain of the text. The two different types of sources are marked up to capture cases where the conceptualizer of the modality is not the producer of the text or speech.

In spoken data, the target is in the scope of an information unit (IU) which may assume different functions: 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) or Locutive Introducer (signals pragmatic suspension of the *hic et nunc* and introduces a

meta-illocution). The BP scheme takes into account, for the annotation of the trigger and the target, the information unit in which they occur: Comment (COM), Topic (TOP), Parenthetical (PAR) or Locutive Introducer (INT). Example (4), taken from Ávila and Melo (2013), illustrates the differences in terms of target delimitation (for an explanation of the transcription symbols, see the authors' paper). The utterance in (4) comprises three different tone units, and the target of the trigger *tem que* 'has to/must', in the second unit, is *restringir também*. It leaves out the direct object of the verb *isso* because it is outside this information unit (defined prosodically). The same sequence in the EP scheme would take as target *restringir também isso*.

- (4) é / [a ɨgente] [tem que]ɨ ɨ[restringir também] / issoɨ //
Yeah / we have to restrict too / this //

The modal cues in both schemes are not restricted to modal auxiliaries, but rather take into consideration a large set of cues, such as propositional verbs, adverbs, adjectives, periphrastic forms and conditionals, and also nouns and interrogative clauses, in EP.

From a preliminary analysis of the two corpora, there are some differences in the set of lexical modal triggers. For instance, a very frequent epistemic trigger in BP is *não tem como*, illustrated in (5), which would translate in EP as *não é possível* and in English as *it is not possible*. There are also less modal adverbs in the corpus of BP than in the EP corpus, and BP favours verbal triggers (97,3%). The two corpora provide data for a future contrastive study of the categories of modal triggers used by both varieties of Portuguese.

- (5) porque eu nunca confundo letras com ɨinformáticaɨ / não
tem nem como //
because I never confuse letters and informatics / it's not
possible //

4 Automatic labeling

The EP corpus of 160K has been the source of data to train an automatic tagger for modality. Many of the verbs in the data set, such as the high frequent modal verbs *querer* 'to want' and *tentar* 'to try', only have one modal value and for those, assigning the correct modal value becomes a trivial task. We focus instead on modal verbs with multiple modal meanings that each occur at least 5 times in the small annotated corpus sample. Only a handful of verbs met this criteria, giving us a

verb	total	modal values		
arriscar	44	epistemic_po: 25	effort: 19	
aspirar	50	volition: 31	epistemic_be: 19	
conseguir	84	success: 41	p-internal_ca: 43	
considerar	29	epistemic_be: 18	evaluation: 11	
dever	108	deontic_ob: 71	epistemic_po: 37	
esperar	52	epistemic_be: 26	volition: 26	
necessitar	50	p-internal_ne: 42	deontic_ob: 8	
permitir	78	epistemic_po: 60	deontic_pe: 18	
poder	236	epistemic_po: 154	deontic_pe: 42	p-internal_ca: 40
precisar	54	p-internal_ne: 45	deontic_ob: 9	
saber	103	epistemic_kn: 93	p-internal_ca: 10	

TABLE 3 Corpus characterization: number of sentences per modal value

list of 11 verbs to work with.

Table 3 shows the 11 verbs and their distribution between the different main modal values (we use abbreviations in the table, for full modal values labels see Table 1). We see that all different modal values of the data set are covered but most verbs only have two different modal values resulting in a sparse matrix. Two of the verbs (*poder* and *dever*) are polysemous verbs that can be used as a semi-auxiliary verb with modal meaning or as a main verb without a modal meaning. We give in (6) an example of a non-modal use of *dever*.

- (6) O clube está em dívida para comigo e o presidente ainda por cima virou-se contra mim a culpar-me de que eu é que *devia* dinheiro ao clube.
 ‘The club is indebted to me and what is more the president turned against me blaming me that I was the one owing money to the club.’

The labeling is done in several steps. First, the dataset is pre-processed with a syntactic parser and the parser output is used to detect the modal triggers. Then, a machine learning approach is used to label each modal trigger with the appropriate modal value in context.

The PALAVRAS parser (Bick, 1999) was used for the syntactic analysis. It distinguishes between modal and non-modal uses of the verbs *poder* and *dever* by labeling the former as auxiliary verbs. By exploiting the parser’s predictions, we were able to detect the modal usage of the two verbs with an F-score of 98%.

Most machine learning algorithms use the vector space model to represent the input data. Using this approach, each sentence is transformed into a set of features that can be boolean, nominal or real valued. Here we use the output of PALAVRAS to build those features: we include information from the trigger itself, from the syntactic tree path and from the trigger's context. We also evaluated two simpler systems, a bag-of-words representation of the sentences and a baseline classifier that always chooses the most common modal value. The next subsection describes in detail the information extracted from the syntactic tree and how it was represented.

The SVM (Support Vector Machine) algorithm (Vapnik, 1998) was chosen to label the modal value of each verb. Several initial experiments were conducted with a bag-of-words representation and with different degrees of the polynomial kernel ($n \in \{1, 2, 3\}$) and values of the C parameter ($C \in \{0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000, 10000\}$). Those experiments enabled us to choose the linear kernel ($n = 1$) with $C = 1$.

Different sets of attributes extracted from the parse tree were evaluated and compared to the bag-of-words approach. For the evaluation we used a 5-fold stratified cross-validation procedure (repeated twice) and computed weighted (macro-)average precision, recall and F1 performance measures. A paired T-test with 95% of significance was applied to analyze the differences to the results of bag-of-words approach. No statistical tests were done over the baseline experiment. These machine learning experiments were conducted using Weka framework (Hall et al., 2009).

4.1 Feature extraction

The feature representation for the machine learning experiments is inspired by the work of Ruppenhofer and Rehbein (2012). Their approach includes three specific sets of attributes, namely:

- from the **trigger**;
- from the **path** of the trigger to the root of the syntactic parse tree;
- from the **context** of the trigger.

Using the PALAVRAS parser, attributes from the trigger, context and path including all possible information given by the parser output were extracted.

We used the following information:

- **trigger**: besides the trigger itself, information about the ancestral nodes (parent, grandparent and great grandparent) was included;

type	source	attributes
trigger	trigger	POS function role morphological semantic
	ancestors	POS function
path	siblings	POS function role morphological semantic
	trigger to root	POS function
context	left/right trigger	POS word lemma

TABLE 4 Attributes extracted from trigger, path and context

- **path**: besides collecting information about the trigger to the root, information about its left and right nodes was also included;
- **context**: using different size windows, information about the previous and following words was included.

These attributes are summarized in Table 4. For each **trigger** word the POS tag, function, morphological and semantic information, and the role (if it exists) were extracted. These tags have the following organization (more detailed information can be found in Bick (2000)):

- 12 POS tags (e.g. nouns, proper nouns, verbs, adverbs, adjectives, pronouns);
- 39 function tags (e.g. subject, accusative, dative, prepositional, adverbial);
- 7 morphological tags (gender, number, case, person, tense, mood, finiteness);
- 171 semantic tags organized in a class taxonomy (e.g. animated human, animated non-human, animated non-human non-moving, topological, actions and events, time field); and

- 45 semantic role labels divided into clause-level (e.g. agent, patient, topic), attributive, spatial, temporal, and adverbial.

For the ancestral nodes (parent, grandparent and great-grandparent) the POS tag and function information was extracted. All attributes are binary representing the presence or absence of that specific value attached to the trigger.

The following information taken from the syntactic parse tree's **path** was extracted: POS tags and functions from the nodes of the path from the trigger to the root and the POS tags, functions, morphological and semantic information, and the role (if exists) of the path of sibling nodes (left and right ones). Counts over each possible value were calculated representing this information as numerical attributes.

The surrounding **context** of each trigger was also considered defining a size window equal to five words (with the trigger word in the middle) and information about the POS tags, words and lemmas was extracted into a set of numerical attributes (representing counts over each possible value).

The number of attributes extracted varies from verb to verb: the minimum number is the set of 552 attributes for the bag-of-words approach of the verb *considerar* and the maximum number is the set of 9770 attributes related to the context of the verb *poder*.

4.2 Experiments and discussion of results

In order to evaluate the discrimination power of each set of attributes, eight experiments were done: bag-of-words, trigger, path, context, trigger+path, trigger+context, path+context and trigger+path+context. Tables 5 and 6 present their precision, and F1 values, respectively; values statistically different (better or worse) from the corresponding bag-of-words experiment are boldfaced.

The results show that the experiments improved the baseline approach for all 11 modal verbs (see tables 5 and 6). Improvements of more than 30 points were obtained for the verbs *arriscar*, *aspirar* and *conseguir*, of around 20 points for the verbs *considerar*, *esperar* and *permitir* and less than 10 points for the verbs *dever* and *saber*. The values for the verb *poder* are lower, but this verb has three modal meanings while the other verbs have two. The improvement is also lower for *necessitar* and *precisar*, but the baseline already achieves high values in these cases, despite very different distributions in terms of frequency of modal value (see table 3).

For the bag-of-words approach, precision values ranging from 0.402 (*considerar*) to 0.815 (*saber*) were obtained; for recall the range was

	base	bow	tg	pth	ct	tg+pth	tg+ct	pth+ct	all
arriscar	.323	.638	.686	.771	.757	.719	.750	.833	.804
aspirar	.384	.741	.853	.795	.694	.778	.756	.778	.779
conseguir	.262	.540	.583	.595	.678	.592	.672	.714	.684
considerar	.385	.402	.489	.526	.611	.582	.536	.650	.660
dever	.432	.700	.662	.568	.626	.636	.692	.611	.602
esperar	.250	.745	.610	.527	.595	.545	.619	.477	.577
necessitar	.706	.708	.723	.735	.732	.709	.701	.690	.698
permitir	.592	.593	.666	.754	.785	.702	.786	.812	.811
poder	.426	.530	.486	.529	.522	.544	.520	.484	.536
precisar	.694	.698	.700	.669	.757	.700	.788	.736	.736
saber	.815	.815	.906	.833	.861	.881	.903	.843	.917

TABLE 5 Results: precision values

between 0.530 (*conseguir*) and 0.903 (*saber*); for the F1 measure values between 0.472 (*conseguir*) and 0.857 (*saber*) were achieved.

	base	bow	tg	pth	ct	tg+pth	tg+ct	pth+ct	all
arriscar	.412	.605	.660	.712	.736	.658	.708	.794	.758
aspirar	.475	.693	.823	.759	.694	.770	.730	.763	.757
conseguir	.347	.552	.554	.581	.670	.563	.640	.689	.650
considerar	.475	.472	.470	.510	.611	.543	.489	.634	.618
dever	.522	.668	.644	.566	.617	.609	.685	.610	.597
esperar	.333	.712	.599	.513	.568	.537	.596	.473	.562
necessitar	.767	.768	.746	.724	.775	.718	.742	.699	.728
permitir	.669	.670	.672	.742	.782	.683	.775	.807	.794
poder	.515	.540	.520	.535	.537	.548	.538	.510	.550
precisar	.758	.760	.716	.697	.787	.721	.799	.763	.770
saber	.857	.857	.895	.849	.884	.889	.914	.872	.923

TABLE 6 Results: F1 values

From Table 5 it is possible to observe that most experiments present no statistical difference to the bag-of-words approach. Nevertheless, when using information extracted from PALAVRAS, an improvement on precision was obtained for the verbs *permitir* and *saber*. For *permitir* better results were obtained using path (0.754), context (0.785), trigger+context (0.786), path+context (0.812) and all (0.811) features; for *saber* better results were obtained with trigger (0.906) and all (0.917) attributes. On the other hand, the verb *esperar* presents worse results when compared to the bag-of-words for path (0.527) experiments.

For recall values, less statistical differences were obtained; only the verb *poder* had different results when compared to the bag-of-words approach (0.650) and they were worse (trigger+path attributes pre-

sented a recall of 0.574 and path+context attributes achieved a recall of 0.551).

Based on Table 6, one can state that using information from the parse tree, improvements on F1 measure were possible for 2 verbs: *permetir* with the path+context setting (0.807) (see section 5) and *saber* with all attributes (0.923).

A follow-up of our experiments will take into consideration ambiguous cases and a new evaluation of the tagging system will be made. Although the 11 verbs were chosen for their polysemy and the ambiguous contexts had been manually tagged with a secondary modal value (in the Ambiguity Attribute of the Trigger component) this information was not taken into consideration in these experiments.

Consider, for instance, the modal verb *dever*. There are 108 modal occurrences of this verb in the corpus, 37 annotated as epistemic possibility and 71 as deontic obligation. In 16 cases of epistemic possibility, the automatic system labeled the context as obligation, while in 17 cases of deontic obligation the system considered the value to be possibility. This leads to a much higher F-measure for the deontic value (0.755) than for the epistemic value (0.521).

However, 5 cases manually tagged as epistemic and automatically labeled as deontic are in fact annotated as ambiguous in the corpus. One such case is example (3), above, where the system predicted only the first reading (epistemic possibility) and ignored the Ambiguity with deontic obligation.

5 Corpus annotation as a source of linguistic insight on modality

The manual annotation of the corpus, its linguistic exploration and the analysis of the results of the automatic labeling have provided valuable insight into the complexity of the semantic concept of modality, its intrinsic ambiguity, the interaction between embedded modal triggers and embedded sources, and the interaction between modality and other systems such as negation. The information gathered during the annotation itself and the error analysis of the tagger provided feedback to improve our modality scheme and is part of the motivation for the presentation of a unifying proposal in section 6. We discuss some aspects where our experiments have proven to shed some light on the concept of modality and its annotation.

- (i) Interaction between modal triggers, negation and focus

Contexts with negative polarity and embedded triggers proved espe-

cially difficult to label with a modal value, even for human annotators. One such example is reproduced in (7): the trigger *deveriam* ‘should’ is preceded by the negation adverb and is followed by another trigger *possibilitado* ‘made possible/enable’. The linguistic context of the trigger is misleading: although there is a negation marker, this operator doesn’t have scope over *deveriam* but instead over its target (it should (not be possible to...)).

- (7) De qualquer forma, depois de falhar a comunicação entre controladores e pilotos, os equipamentos electrónicos não **deveriam** igualmente ter possibilitado que se chegasse a um ponto em que um dos aviões tivesse que fazer uma descida tão abrupta como a que fez o avião da TAP.
 ‘Anyway, after the communication between controllers and pilots failed, the electronic equipment should not have let things get to a point where one of the planes took such an abrupt descent such as TAP airplane did.’

Example sentences (8) and (9) show two embedded triggers that both have negative polarity values. In sentence (8), the deontic verb *poder* ‘should’ has negative polarity and the embedded trigger *conceder* ‘to concede’ is preceded by the aspectual expression *deixar de* ‘to cease’ that carries intrinsic negative polarity. The double negation (should not cease to concede) is equivalent to a global positive polarity (should concede). In (9), the first trigger conveys an epistemic value and a negative polarity (*não julgo* ‘[I] don’t believe’), while its target contains a second trigger that lexically expresses negative polarity (*imprescindível* ‘indispensable’) and denotes a deontic obligation value. The overall interpretation of the sentence is that the source believes that something is dispensable. In both cases, our annotation treats each trigger independently and does not express the overall positive polarity of the clause. The modal interpretation of an embedded trigger may also be affected by the preceding trigger. In sentence (10), the trigger *pode* ‘can’ influences the degree of certainty of the modal value evaluation of the second trigger *difícil* ‘difficult’.

- (8) Não se pode, aliás, deixar de conceder cabimento à formulação (...).
 ‘We shouldn’t, in fact, cease to concede authorisation to the drafting (...).’

- (9) Não julgo imprescindível que se encontre desde já, a correr, o novo rosto.
‘I don’t think that it is indispensable that we find right now, immediately, the new face’.
- (10) Se o aluno se perde, pode ser difícil voltar a apanhar.
‘If the student is lost, it can be difficult for him to catch up again’.

Despite the challenge posed by these contexts, we believe that it is important to individually mark each element of the sentence that contributes to the modal values to be able to understand how each provides input to the overall semantic interpretation. The global interpretation of the sentence does not only depend on the modal information but is determined by an interplay of information at different levels including focus and negation. Investigating these interactions at the global sentence level is beyond the scope of this current work but promises to be a fascinating future research direction.

The annotation of corpus data also points to the fact that exclusive particles, such as *só* ‘only’, may affect and alter the modal meaning of the sentence. For instance, in contexts where the verb *poder* has an epistemic reading, adding the exclusive can restrict the set of possibilities to the possibility presented in the sentence, as illustrated in (11).

- (11) Isto só pode ter sido um acidente.
‘This can only have been an accident’

In (11), by restricting the set of possible situations to one (x and only x), the adverb leads to an overall reading of the sentence that expresses epistemic necessity. This does not hold, however, in all such contexts: the interpretation requires that the target of the modal trigger is a state or a past event, and that the exclusive particle has scope over the full target of the modal trigger. Compare, for instance, the modal readings of the corpus sentence (12) and the slightly adapted version in (13). In (12), the exclusive has scope over the temporal adjunct and denotes a condition over a possibility, while in (13) it has scope over the full target and the overall reading is one of epistemic necessity.

- (12) Ora, a Sr.^a Deputada MFL *só pode ter razão quando acertar nalguma previsão.*
 ‘Well, the member of parliament MFL may only be right when at least one of her forecasts turns out correct.’
- (13) Ora, a Sr.^a Deputada MFL *só pode ter razão.*
 ‘Well, the member of parliament MFL can only / must be right.’

The exclusive particle can also mark the possibility as weaker than expected. Exclusives can downtone an alternative, by underlining the fact that this proposition “is not the strongest that in principle might have been the case”, a function called *Mirative* (Beaver, 2008). In (14), the combination of the exclusive with the deontic marker doesn’t express one obligation among a set of possibilities, but rather states that the alternative is much easier than one would expect.

- (14) Para participar só tem de contactar a organização através dos telefones 96... ou 91...
 ‘To participate, you only have to contact the organization through the phone numbers...’

These meaning interactions are extremely challenging for any annotation scheme and these contexts tested the limits of what could be represented in our current schema. To represent and compute the meaning of such sentences, negation and focus (and any other element that affects modality) could either: (i) be integrated in the modality scheme, an approach that we follow partly by marking the polarity of trigger and target, and also by proposing to deal with exclusive particles as another modal marker in the modal set (Mendes et al., 2013); (ii) be the topic of separate layers of annotation with their own annotation scheme; (iii) be included as features of a global annotation scheme that would take the event as the tagging unit, instead of individual triggers.

(ii) The attributes *Path* and *Context*

The discussion of the experiments with different sets of attributes in section 4 pointed out that the system achieved statistically significant improvement over the bag-of-words approach with the path+context settings. By running an attribute ranker (using information gain), we were able to single out the most informative attributes for *path*:

- presence of an Accusative node between the root and the verb node
- no explicit subject in the left brother node

- the left brother node receives the semantic role Theme
- presence of an infinitive clause in the path: either from the root to the verb or as the right brother node
- the left brother node is a Dative object with the function Beneficiary

For *context*, the more important attributes occur in the left tree (we single out the ones that complement the information in *path*):

- the lemma *lei* ‘law’ occurs in the left context
- the dative clitic *lhe* ‘to him/her’ occurs in the left context

The combination of these attributes express certain properties that seem to favour one modal value over another. For instance, the presence of an Accusative node of the type infinitival clause favours an epistemic reading, as illustrated in (15). There are 77 contexts with an Accusative infinitival clause with the verb *permitir*: 4 of these contexts have a deontic reading (about a quarter of the total number of deontic contexts with *permitir*), while 37 have an epistemic reading (about half of the frequency of epistemic contexts with this verb). Moreover, many of the examples of epistemic possibility reading with *permitir*, such as (16), are associated to constructions where the left brother node is a Dative object: the 10 cases of Dative occurring in pre-verbal position are all instances of epistemic possibility.

- (15) Mas estes primeiros dias já *permitem* tirar conclusões.
 ‘But these first days already make it possible to draw conclusions.’
- (16) Agora, embora não seja capaz de pintar porque não tenho técnica para o fazer, descobri que o computador *me permite* transformar as minhas imagens de tal maneira que ficam a parecer autênticas pinturas.
 ‘Now, although I’m not capable of painting because I don’t have the technique to do so, I discovered that the computer allows me to transform my images in such a way that they end up looking like authentic paintings.’

A non-explicit subject in the left brother node is, on the contrary, associated to a deontic reading of *permitir*. In fact, all deontic contexts of this verb occur in subordinated clauses (mostly relative clauses, as illustrated in (17)) or coordinated clauses. The presence of the lemma *lei* ‘law’ in the left context, specifically in subject position of *permitir*, is a clear indication of a deontic obligation reading (there are 5 occurrences of a subject containing the lemma *lei*, where 4 are deontic), as in (18). In fact, this annotation is in accordance to the fact that deontic

necessity derives from some source or cause which may be a person, institution or *more or less explicitly formulated body of moral or legal principles* (John Lyons, 1977, pp. 824)

- (17) São contratos em condições *suis generis*, *que permitem* prazos de pagamentos de oito e mais anos (...).
 ‘These are special contracts, that allow payment deadlines of eight years and more (...).’
- (18) E acrescenta que não existe nenhuma lei que permita à Portugal Telecom cortar o serviço telefónico por os utentes não pagarem, por exemplo, as chamadas de valor acrescentado, tipo telefonemas eróticos, etc.
 ‘And [he/she] adds that there is no law that allows Portugal Telecom to cut the phone service when users don’t pay, for instance, value added calls, such as erotic phone calls.’

The role played by some of these attributes in the labeling of *permitir* is quite unexpected and was uncovered by the analysis of the results of the system. This was the case, for instance, of the attributes “no explicit subject” and “Dative brother node in left context”.

(iii) Ambiguity

Our experiment focused on attributing the correct modal value to modal triggers that may denote more than one modality type. In most cases, the context provides information for disambiguation and a single value was attributed by our human annotators. The goal is for the system to also decide, based on the context, which modal value applies to the sentence. As we mentioned in section 3.1, the context might not provide enough information for manual disambiguation and the human annotators labeled these cases as ambiguous. In future experiments, we will take these cases into consideration and the outcome of the automatic labeling should be able to mark these sentences with two potential modal meanings.

Ambiguity has typically no formal correlate in the nature and span of the components Trigger and Target of our annotation. For instance, in example (19), *permitir* has, on the one hand, an epistemic possibility reading, in the interpretation that the climate makes it possible for trees to grow. On the other hand, it expresses deontic permission if we consider that the climate is a necessary condition for the growth of the trees. In both cases, the subject NP *as condições climáticas* will be marked as the source (of possibility or necessity) and the object will be the target.

- (19) 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)’.

However, there might be a formal counterpart to semantic ambiguity. It is the case of contexts that are ambiguous between deontic obligation and participant-internal necessity with verbs *dever* and *ter de*, and epistemic possibility and participant-internal capacity with verb *poder*. In (20), two interpretations - and therefore two annotations - are possible. The epistemic possibility reading might be paraphrased as *it is possible that [he is an excellent candidate]*, so that the target of the modal trigger is the whole proposition and includes the subject of the sentence *Ele@ser um excelente candidato*⁴. The source of the modality will then be the speaker/writer. The internal capacity reading can be paraphrased as *he has the capacity to be an excellent candidate*: the source of the capacity is the subject and the target is the verbal phrase. Currently the scheme has no way of capturing these structural ambiguities and only one structure, representing what the annotator considers to be the most salient meaning in the context, is expressed, although semantic ambiguity is marked in the Ambiguity attribute of the trigger. Accordingly, we expect our system to be able, in future experiments, to label these contexts with two possible modal meanings.

- (20) Ele pode ser um excelente candidato.
 ‘He might be an excellent candidate’
Modal value: epistemic possibility
 Target: Ele@ser um excelente candidato
 Source of the event mention: sp/wr
 Source of the modality: sp/wr
Modal value: participant-internal capacity
 Target: ser um excelente candidato
 Source of the event mention: sp/wr
 Source of the modality: Ele

The ambiguous modal triggers that we have discussed so far can denote two or more individual modal meanings and, in most cases, the context provides the necessary information to select the appropriate meaning (see Kratzer (1991) and Coates (1983) for different concepts of

⁴Note that the discontinuity of the target is marked here with the symbol @, but is encoded in XML in the data set.

the polysemy vs. indeterminacy of modal verbs). In these cases, only one reading is available at a time and the interpretations belong to different modal categories (e.g. epistemic vs. deontic) (Coates, 1983). However, the verb *conseguir* ‘to succeed’ seems to behave differently in that the capacity reading and the success reading are not non-compatible discrete meanings, but rather intertwined aspects of the semantics of the verb. In fact, in most contexts of the verb, the value success is associated with an internal capacity of the subject. This type of semantic indeterminacy is related to the concept of gradience in the work of Coates (1983) and would differ from cases of pure ambiguity.

This is particularly important to assess, first, the relations that exist between the modal values included in the scheme to address any possible overlap in the case of some lexical triggers (not always foreseeable when establishing which modal values to include), and second the performance of the annotation tool. Cases of gradience, where two or more values are simultaneously present in a single context, although one might be more salient than the other, will be especially difficult to tag by a manual annotator (the choice might be more prone to context influence, such as perfective tenses leaning to the success reading due to their association to factuality).

6 Future work: An unified scheme for Portuguese and its application

The proliferation of annotation schemes for modality is inevitable and the result of specific objectives of the different teams working on the topic. However, some attempt of standardization would be of interest to the field, making contrastive studies an attainable goal. In the case of the EP and BP schemes, the objectives are quite similar and the properties of both varieties do not differ in the components of the schemes, although the list of lexical triggers might be variety-specific to a certain extent. The preparation of a unified proposal is also an opportunity to refine both schemes according to our experience and the knowledge acquired during the manual and automatic labeling.

As mentioned in section 3, the set of components is practically identical in each scheme. The differences arise essentially in the list of modal values, the Target dependent component and the trigger and target attributes. Let us start with the mismatches in modal values. Table 7 presents a comparison of the modal values that are considered in the EP and the BP modal schemes: equivalent modal values (or sub-values) are presented in the same row, regardless of their designation. Both schemes are organized in terms of main and secondary modal values.

Most of the modal values are included in both schemes: epistemic possibility, epistemic knowledge, epistemic belief, deontic obligation, deontic permission, capacity/ability, volition. There are some mismatches: the contexts tagged with the sub-value epistemic necessity in BP seem to be close to the value deontic obligation in EP (we mark these cases between // in Table 7); the deontic prohibition value in BP is most probably annotated as a deontic permission with negative polarity in EP; and participant-internal necessity in EP is covered by deontic necessity in BP. Two sub-values have no equivalent: epistemic probability only occurs in BP and epistemic interrogative only occurs in EP. Besides those sub-values, three main values in the EP scheme are absent in BP: evaluation, effort and success. There is however a partial equivalence for Success: when success is related to an internal capacity (e.g. verb *consequir* ‘achieve’) it is tagged as dynamic ability in BP.

EP modal scheme	BP modal scheme
<i>Epistemic</i>	<i>Epistemic</i>
Possibility	Possibility
	Probability
/Deontic/	Necessity
Knowledge	Knowledge
Belief	Belief
Doubt	Verification
Interrogative	
<i>Deontic</i>	<i>Deontic</i>
Obligation	Obligation
Permission	Permission
/Deontic perm., neg. polarity/	Prohibition
/Internal necessity/	Necessity
<i>Participant internal</i>	<i>Dynamic</i>
Necessity	/Deontic necessity/
Capacity	Ability
Volition	Volition/Intention
Evaluation	
Effort	/Dynamic ability/
Success	/Dynamic ability/

TABLE 7 Comparison of modal values in the EP and BP schemes

We present our proposal for a unifying set of categories in Table 8. Although the percentage of occurrence of the epistemic probability value is relatively low in the BP corpus, this value is nevertheless

important in the modality typology and quite easily distinguishable from epistemic possibility. These two subvalues of epistemic modality are covered by the pairs of lexical items *poder/dever* ‘might/should’, *possível/provável* ‘possible/probable’, *possibilidade/probabilidade* ‘possibility/probability’. Consequently, we keep this value in the final set. The uncertainty meaning conveyed by the epistemic verification value (BP) is in fact covered by the more general epistemic possibility value. The same is valid for epistemic doubt (EP), which translates into an epistemic possibility value with negative polarity (I doubt that this will happen = maybe it is not possible that this will happen). Direct interrogative sentences are syntactically marked as such and their annotation as modal instances in the EP scheme involved marking the entire sentence as trigger and target, what seems unnecessary. Indirect interrogative sentences express a possibility value that can be captured as such in the scheme. Necessity is a concept that required further revision in both schemes: the EP scheme doesn’t capture contexts where necessity is the result of circumstances (Circumstantial modality or participant-external modality). In spite of the difficulty in establishing whether a necessity is external or instead is an obligation established by the entities involved in the state of affairs, it is important to make the distinction in the clear-cut cases. With this in mind, we keep the value deontic necessity (*é necessário que* ‘it is necessary that’). We also keep the dynamic value (BP) instead of the participant-internal one (EP). However, we enlarge the sub-values of the dynamic category, so as to include several categories related to the expression of a subjective attitude of the subject. It is the case of necessity, ability and volition. Since effort and success are types associated to the dynamic ability sub-value, we decided to leave them out. Finally, we keep the category evaluation as crucial for studies of belief and opinion, and we would like to study this category in more detail in the future.

In the Target component, the difference between the two schemes lies in the type of segment which is tagged in the corpus: a syntactic phrase or any locutory material in the scope of an information unit. We think that the functions of the information units should be the subject of a separate layer of annotation: the information structure. The Target dependent component should be addressed in the co-reference level of annotation. Therefore, we keep the single Target component in the unified scheme. In Table 8, we presented a unified annotation scheme with a list of components, the attributes of the trigger and the list of modal values that is applicable to both spoken and written, European and Brazilian, Portuguese.

Components	Attributes	
Trigger	Polarity	
	Ambiguity	
	Modal type	
	Modal values	Modal sub-values
	Epistemic	Possibility; Probability; Knowledge; Belief
	Deontic	Obligation; Permission; Necessity
	Dynamic	Necessity; Ability; Volition; Evaluation
Target	Polarity	
Source of the modality		
Source of the event mention		

TABLE 8 Unified schema proposal

7 Conclusion

We have presented two modality schemes for two varieties of Portuguese, covering written and spoken registers, and an experiment to automatically annotate modality for 11 Portuguese modal verbs, using a corpus sample of 160K tokens, manually tagged with modal values.

The range of phenomena that are not considered as modality in EP and BP schemes are very similar, and so are the components of both modality schemes. Differences are essentially genre-related since the BP scheme is applied to spontaneous speech and considers the scope of the modality to be the information unit. The EP scheme was applied to a sample from the EP written sub-part of the Reference Corpus of Contemporary Portuguese (CRPC), while the BP scheme was applied over a sample from the C-ORAL-BRASIL. Based on our annotation experience of these two corpus samples and on the automatic experiments on the EP corpus, we prepared a unified proposal for modality annotation in Portuguese that applies to both written and spoken modality.

For the automatic labeling of European Portuguese, we selected verbs that had more than one modal meaning and occurred at least 5 times in the corpus sample. Using the Weka framework we conducted several experiments to study the effect of the usage of linguistic information to identify modal values. We compared those against the majority baseline and a bag-of-words approach and calculated precision, recall and F_1 measures. The system performed well above the majority baseline. We were able to get better results (precision and F1) with some settings for some verbs (*permitir* and *saber*), but most experiments, even with higher performance values, were not significantly different from the bag-of-words approach. We assume that this was mainly due to the small number of training examples. We studied a range of au-

tomatically generated linguistic attributes that can be used to identify the modal values for Portuguese verbs. We can conclude from these initial experiments that the use of information extracted from parse trees does not harm the performance of the automatic taggers and can, for some verbs and combinations, enhance it. Considering that our training corpus was relatively small and that we selected challenging verbs in our experiment, we believe that our goal, of creating a larger corpus with modal information by a (semi) automatic tagging process, could lead to positive results in the future.

As a next step we aim to run new experiments with the modality tagger using the unified scheme as proposed in section 6 and apply the tagger both to European written and Brazilian spoken Portuguese data. The current modality tagger uses output from the PALAVRAS parser to create its feature representation. The PALAVRAS parser has been recently adapted to run on the spoken corpus CORAL-BRASIL (Bick et al., 2012) and we can expect the parser to work well on the two different textual genres.

We plan to follow up this analysis with a detailed study identifying the individual role of the syntactic and semantic features that are used for the automatic attribution of the modal value in our system. More specifically, for each parse tree we intend to evaluate the relevance of the following features in the modal value classification: word, lemma, POS tag, syntactic tag, semantic information and role label. The analysis will be performed over the partial parse trees that include the modal verbs and their parents and grand-parents and also over the nodes in the path from the root of the sentence parse trees to the modal verbs. We will also make a comparative study of the relevant features for each of the studied modal verbs. In fact, from the analysis of the results that we obtained so far, we suspect that there are important differences between different modal verbs. The fact that, for some verbs, the parse tree input obtains better results and for others it is better to use the list of words of the sentences, suggests that syntactic and semantic features might not be equally relevant for all verbs.

We also aim to compute a learning curve to estimate the amount of manually annotated examples that are needed to get a good performance from the modality tagger. Furthermore, to be able to label new verbs that did not occur in the initial data set, we plan to train a general modal trigger classifier that is not dependent on the verb itself.

Acknowledgments

This work was partially supported by national funds through FCT - Fundação para a Ciência e Tecnologia, under project Pest-OE/EEI/LA0021/2013 and project PEst-OE/LIN/UI0214/2013, and through FAPEMIG (PEE-00293-15).

References

- Ávila, Luciana and Heliana Melo. 2013. Challenges in modality annotation in a Brazilian Portuguese Spontaneous Speech Corpus. In *Proceedings of IWCS 2013 WAMM Workshop on the Annotation of Modal Meaning in Natural Language*. Potsdam, Germany: Association for Computational Linguistics.
- Ávila, Luciana Beatriz. 2014. *Modalidade em perspectiva: estudo baseado em corpus oral do Português Brasileiro*. Tese de Doutorado, Universidade Federal de Minas Gerais, Belo Horizonte.
- Baker, Kathrin, Michael Bloodgood, Bonnie Dorr, Nathaniel W. Filardo, Lori Levin, and Christine Piatko. 2010. A modality lexicon and its use in automatic tagging. In N. Calzolari, C. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias, eds., *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. Valletta, Malta: European Language Resources Association (ELRA). ISBN 2-9517408-6-7.
- Bally, Charles. 1932. *Linguistique générale et linguistique française*. Berna: Francke Verlag.
- Beaver, David I. 2008. *Sense and Sensitivity: How Focus Determines Meaning*. Blackwell Pub.
- Bick, Eckhard. 1999. *The parsing system PALAVRAS*. Aarhus University Press.
- Bick, Eckhard. 2000. *The Parsing System "Palavras". Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. \Aarhus: University of Arhus.
- Bick, Eckhard, Heliana Mello, Alessandro Panunzi, and Tommaso Raso. 2012. The annotation of the c-oral-brasil spoken corpus using an adaptation of the palavras parser. In N. Calzolari, K. Choukri, T. Declerck, M. U. Dogan, B. Maegaard, J. Mariani, J. Odijk, and S. Piperidis, eds., *LREC'2012 – Eighth International Conference on Language Resources and Evaluation*, pages 3382–3386. European Language Resources Association (ELRA). ISBN 978-2-9517408-7-7.
- Bybee, Joan and Suzanne Fleischman. 2013. *Modality and grammar in discourse*. Amsterdam/Philadelphia: John Benjamins.
- Coates, Jennifer. 1983. *The semantics of modal auxiliaries*. London, Canberra: Croom Helm.
- Cohen, Jacob. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement* 20(1):37.

- Cresti, Emanuela. 2000. *Corpus di italiano parlato*. Firenze: Accademia della Crusca.
- Cresti, Emanuela and Massimo Moneglia. 2005. *C-ORAL ROM: Integrated reference corpora for spoken Romance languages*. Amsterdam/Philadelphia: John Benjamins.
- Cresti, Emanuela and Antonietta Scarano. 1998. *Sur la notion de parlé spontané*.
- Diab, Mona T., Lori S. Levin, Teruko Mitamura, Owen Rambow, Vinodkumar Prabhakaran, and Weiwei Guo. 2009. Committed belief annotation and tagging. In *Third Linguistic Annotation Workshop*, pages 68–73. Singapore: The Association for Computer Linguistics. ISBN 978-1-932432-52-7.
- Farkas, Richárd, 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. Uppsala, Sweden: Association for Computational Linguistics.
- Généreux, Michel, Iris Hendrickx, , and Amália Mendes. 2012. Introducing the reference corpus of contemporary portuguese on-line. In N. Calzolari, K. Choukri, T. Declerck, M. U. Dogan, B. Maegaard, J. Mariani, J. Odiijk, and S. Piperidis, eds., *LREC'2012 – Eighth International Conference on Language Resources and Evaluation*, pages 2237–2244. Istanbul, Turkey: European Language Resources Association (ELRA).
- Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The weka data mining software: An update. *SIGKDD Explor. Newsl.* 11(1):10–18.
- Hendrickx, Iris, Amália Mendes, and Silvia Mencarelli. 2012a. Modality in text: a proposal for corpus annotation. In N. Calzolari, K. Choukri, T. Declerck, M. U. Dogan, B. Maegaard, J. Mariani, J. Odiijk, and S. Piperidis, eds., *LREC'2012 – Eighth International Conference on Language Resources and Evaluation*, pages 1805–1812. Istanbul, Turkey: European Language Resources Association (ELRA). ISBN 978-2-9517408-7-7.
- Hendrickx, Iris, Amália Mendes, Silvia Mencarelli, and Agostinho Salgueiro. 2012b. *Modality Annotation Manual*. Centro de Linguística da Universidade de Lisboa, Lisboa, Portugal, v1 edn.
- John Lyons. 1977. *Semantics*, vol. 2. Cambridge: Cambridge University Press.
- Kratzer, Angelika. 1991. Modality. In Arnim v. Stechow and Dieter Wunderlich, eds., *Handbuch Semantik/Handbook Semantics*, pages 639–650. Berlin, New York: de Gruyter.
- Kratzer, Angelika. 2013. *Modals and Conditionals: New and Revised Perspectives*. Oxford: Oxford University Press.
- Lee, Kenton, Yoav Artzi, Yejin Choi, and Luke Zettlemoyer. 2015. Event detection and factuality assessment with non-expert supervision. In

- L. Màrquez, C. Callison-Burch, J. Su, D. Pighin, and Y. Marton, eds., *Proceedings of the 2015 Conference on Empirical Methods on Natural Language Processing*, pages 1643–1648. The Association for Computational Linguistics.
- Martin, Pierre. 2004. Winpitch corpus, a text to speech alignment tool for multimodal corpora. In *Proceedings of the Fourth Conference on the International Language Resources and Evaluation (LREC'04)*. Lisbon, Portugal.
- Matsuyoshi, Suguru, Megumi Eguchi, Chitose Sao, Koji Murakami, Kentaro Inui, and Yuji Matsumoto. 2010. Annotating event mentions in text with modality, focus, and source information. In N. C. C. Chair), K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias, eds., *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. Valletta, Malta: European Language Resources Association (ELRA). ISBN 2-9517408-6-7.
- McShane, Marjorie, 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. Ann Arbor, Michigan: Association for Computational Linguistics.
- Mello, Heliana and Tommaso Raso. 2011. Illocution, modality, attitude: different names for different categories. In H. Mello, A. Panunzi, and R. T., eds., *Pragmatics and prosody: illocution, modality, attitude, information patterning and speech annotation*, pages 1–18. Firenze: Firenze University Press.
- Mendes, Amália, Iris Hendrickx, Agostinho Salgueiro, and Luciana Ávila. 2013. Annotating the interaction between focus and modality: the case of exclusive particles. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 228–237. Sofia, Bulgaria: Association for Computational Linguistics.
- Miwa, Makoto, Paul Thompson, John McNaught, Douglas B. Kell, and Sophia Ananiadou. 2012. Extracting semantically enriched events from biomedical literature. *BMC Bioinformatics* 13:108.
- Morante, Roser and Caroline Sporleder. 2012. Modality and negation: An introduction to the special issue. *Computational Linguistics* 38(2):223–260.
- Moreira, Benjamim. 2005. *Estudo de alguns marcadores enunciativos do português*. PhD Dissertation, Universidade de Santiago de Compostela, Faculdade de Filologia, Santiago de Compostela, Spain.
- Müller, Christoph and Michael Strube. 2006. Multi-level annotation of linguistic data with mmax2. *Corpus technology and language pedagogy: New resources, new tools, new methods* 3:197–214.
- Nirenburg, Sergei and Marjorie McShane. 2008. Annotating modality. Tech. rep., University of Maryland, Baltimore County, USA.

- Nissim, Malvina and Paola Pietandrea. 2015. Preface. In *In Proceedings of the IWCS Workshop on Models for Modality Annotation, MOMA 2015*. London: The Association for Computational Linguistics.
- Nissim, Malvina, Paola Pietrandrea, Andrea Sanso, and Caterina Mauri. 2013. Cross-linguistic annotation of modality: a data-driven hierarchical model. In *Proceedings of the 9th Joint ISO - ACL SIGSEM Workshop on Interoperable Semantic Annotation*, pages 7–14. Potsdam, Germany: Association for Computational Linguistics.
- Oliveira, Fátima. 1988. *Para uma semântica e pragmática de DEVER e PODER*. Dissertação de Doutoramento em Linguística Portuguesa, Universidade do Porto, Faculdade de Letras, Porto.
- Palmer, Frank R. 1986. *Mood and Modality*. Cambridge textbooks in linguistics. Cambridge University Press.
- Palmer, Frank R. 2001. *Mood and Modality*. Cambridge textbooks in linguistics. Cambridge: Cambridge University Press, v2 edn.
- Prabhakaran, Vinodkumar, Michael Bloodgood, Mona Diab, Bonnie Dorr, Lori Levin, Christine D Piatko, Owen Rambow, and Benjamin Van Durme. 2012. Statistical modality tagging from rule-based annotations and crowdsourcing. In *Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics*, pages 57–64. Association for Computational Linguistics.
- Raso, Tommaso and Heliana Mello. 2012. *C-ORAL-BRASIL I: Corpus de referência do português brasileiro falado informal e DVD multimedia*, vol. I. Belo Horizonte: Editora UFMG.
- Rubinstein, Aynat, Hillary Harner, Elizabeth Krawczyk, Daniel Simonson, Graham Katz, and Paul Portner. 2013. Toward Fine-grained Annotation of Modality in Text. In *Proceedings of the Tenth International Conference for Computational Semantics (IWCS 2013)*.
- Ruppenhofer, Josef and Ines Rehbein. 2012. Yes we can!? Annotating English modal verbs. In N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, J. Odijk, and S. Piperidis, eds., *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey: European Language Resources Association (ELRA). ISBN 978-2-9517408-7-7.
- Sauri, Roser, Marc Verhagen, and James Pustejovsky. 2006. Annotating and Recognizing Event Modality in Text. In *FLAIRS Conference*, pages 333–339.
- Song, Zhiyi, Ann Bies, Stephanie Strassel, Tom Riese, Justin Mott, Joe Ellis, Jonathan Wright, Seth Kulick, Neville Ryant, and Xiaoyi Ma. 2015. From light to rich ere: Annotation of entities, relations, and events. In *Proceedings of the 3rd Workshop on EVENTS at the NAACL-HLT 2015*, pages 89–98. The Association for Computational Linguistics.
- Szarvas, György, Marc Verhagen, Richárd Farkas, G. Mora, and János Csirik. 2008. The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics* 9 (Suppl 11)(S9).

- Tucci, Ida. 2008. La modalità nel parlato spontaneo e il suo dominio de pertinenza. una ricerca corpus-based (C-ORAL-ROM Italiano). In *Actes du XXVe CILPR*.
- van der Auwera, Johan and Vladimir A. Plungian. 1998. Modality's semantic map. *Linguistic Typology* 2:79–124.
- Vapnik, Vladimir N. 1998. *Statistical Learning Theory*. Wiley-Interscience.