Cross-linguistic annotation of modality: a data-driven hierarchical model

Malvina Nissim
University of Bologna

Paola Pietrandrea
Lattice-CNRS, France

Andrea Sansò
University of Insubria

Caterina Mauri
University of Pavia

Abstract

We present an annotation model of modality which is (i) cross-linguistic, relying on a wide, strongly typologically motivated approach, and (ii) hierarchical and layered, accounting for both *factuality* and *speaker's attitude*, while modelling these two aspects through separate annotation schemes. Modality is defined through cross-linguistic categories, but the classification of actual linguistic expressions is language-specific. This makes our annotation model a powerful tool for investigating linguistic diversity in the field of modality on the basis of real language data, being thus also useful from the perspective of machine translation systems.

1 Introduction and Background

A text cannot be simply regarded as a sequence of representations of State of Affairs (SoAs) occurring (or having occurred) in the actual world. Texts may comprise representations of counterfactual or non factual SoAs, as well as a number of expressions encoding the stance the writer/speaker might be taking on a SoA, implying different attitudes, possibly relying on external sources of information. These aspects fall under the more general label of *modality*.

The automatic interpretation of modality can be seen as two tasks: (i) identifying the representations that are not put forward as factual and (ii) identifying the sentiments or opinions speakers may have towards their representations. These two tasks, which we call *factuality mining* and *speaker's attitude mining*, respectively, are two independent, albeit often related, semantic and linguistic dimensions.

Since the first step towards developing systems which deal with modality automatically is the creation of appropriate, annotated resources, the last few years have witnessed the development of annotation schemes and annotated corpora for different aspects of modality in different languages (McShane et al. (2004); Wiebe et al. (2005); Szarvas et al. (2008); Sauri and Pustejovsky (2009); Hendrickx et al. (2012); Baker et al. (2012)).

While important contributions, these remain mainly separate efforts. And while there have been efforts towards finding a common avenue for modality annotation, such as the CoNLL-2010 Shared Task, ACL thematic workshops and a special issue of Computational Linguistics (Morante and Sporleder (2012)), the computational linguistics community is still far from having developed working, shared standards for converting modality-related issues into annotation categories.

Linguistic theory, and especially linguistic typology, has already gone a long way in the study of modality across languages. However, this very aspect of cross-linguality has been overlooked in devising annotation schemes. Instead, we believe that working in a multilingual environment could ease the annotation, and at the same time make it more semantically meaningful, by keeping the layer of functional categories distinct from their actual linguistic realisations. Indeed, modality can be modelled more elegantly and efficiently starting from a functional, higher, level, while languages encode with their own means the specified concepts and categories.

Therefore, we promote an annotation model of modality which is (i) cross-linguistic, relying on

a wide, strongly typologically motivated approach, (ii) adaptable, capable of accounting for the linguistic realisation of modality in each single language under consideration; and (iii) hierarchical and layered, accounting for both *factuality* and *speaker's attitude*, while modelling these two aspects through separate annotation schemes. Within this frame, the issue of *annotation units*, linguistically, becomes crucial, and we claim that such a two-layered framework provides the best setting for dealing with it.

2 Annotating Modality

In spite of the large amount of solid work on modality in theoretical linguistics and linguistic typology, and in spite of the various more NLP-oriented annotation schemes that are flourishing in the last years, there are as yet no shared standards for modality annotation. This is extreme to the point that Vincze et al. (2010) have observed, through a very detailed analysis and classification of problematic issues, that the same biomedical data was annotated in two different projects yielding minimal overlap, both semantically and syntactically.

A main issue is that there is no actual consensus on the very **notion of modality** to be translated into annotation categories. While it is *factuality* the key notion in Sauri and Pustejovsky (2009)'s FactBank, for instance, it is instead the *speaker's attitude* that is addressed in other recent annotation exercises (Nirenburg and McShane (2008); Hendrickx et al. (2012); Baker et al. (2012)).

Also not uniform across different projects is the actual **annotation procedure**, in terms of which functional categories must be annotated in text. It is quite common to consider the trigger, the scope, and the source (or author) as relevant categories, but not all of them translate into actual annotation. For example, in (Baker et al. (2010)), all three of them are signalled to the annotators in text, but it is only the targets which are to receive an annotation value.

And crucially, there are wide differences, and often little clarity, in terms of which **linguistic units** should be annotated. It has been shown in typological and constructional studies on modality that modal triggers may vary in nature and complexity (morphemes, verbs, adverbs, complex constructions, etc.) and that the scope of a modal marker may vary in extension from a single word to an entire text Masini and Pietrandrea (2010). One major problem is that in a few projects the annotators are not asked to select the annotation units but only to assign modality values to preselected markables, thus turning annotation into a classification task. In their annotation guidelines, Baker et al. (2010) assume that the units to be marked up are already highlighted and do not exceed the clause limit (i.e. the maximum extension is a phrase) and revolve around a verb, but it isn't clearly specified how such units are selected, nor why. Differently, Hendrickx et al. (2012) let the annotators choose the unit and its extension, allowing also for cross-sentential markables to be selected. However, they pre-select data to be annotated by matching a finite set of modality trigging verbs, thereby also imposing some degree of constraint. While pre-selecting annotation units maximises homogeinity and reduces disagreement among annotators, it is not clear exactly *which* units are to be marked up and whether it is at all an appropriate procedure in all cases.

Building on insights coming from linguistic typology, we will take a stand on these issues and claim that a cross-linguistic perspective provides the best framework for devising an annotation model for modality. We will also claim that the issue of annotation units must be addressed, and it becomes more meaningful and better dealt with within such a framework, thanks to a division between a *functional annotation*, where functional categories are specified and a *linguistic annotation* where actual units are selected for annotation, depending on the language. In Section 5 we will show how we suggest to combine these two different annotations.

3 A two-layered approach

Two related but distinct phenomena are often lumped together under the label of modality: factuality and speaker's attitude.

Factuality A representation can be put forward as depicting an event actually occurring or having occurred (factual SoAs, 1a), an event having not occurred in the real world (counterfactual SoAs, 1b), or not grounded in reality (non factual SoAs, 1c):

(1) a. He came

b. He did not come

c. She fears he came

As the examples show, the representation as such does not encode the factuality of the depicted event. It is only the context that allows for a specification of this value.

Speaker's attitude Speaker's attitude may also contribute to specify the factuality of a SoAs, but it does so only incidentally. The main purpose of the markers of speaker's attitude is specifying the stance of the speaker towards his representation, rather than the factuality status of that representation. The speaker can express his commitment about the SoA (epistemic modality), whether expressing his genuine commitment (commitment) or specifying the evidence he has for his opinion (evidential epistemic); he can can manifests his will concerning the SoA (deontic modality), whether expressing a mere wish (volitional deontic) or manipulating the adressee toward the realisation of the SoA (manipulative deontic); he can express his moral or esthetic judgment about a SoA or his fear about it (evaluative modality).

Two orthogonal dimensions Sometimes the expression of a given speaker's attitude entails the non factuality of a representation (2a), but this is not always the case (2b)

(2) a. I am afraid that he does not misses me

b. It's scary that he does not misses me

On the other hand the non-factuality of a representation may be encoded by means other than speaker's attitude markers, such as hypothetical subordinating conjunctions (3a), or alternative constructions (3b):

(3) a. if he misses me, I am happy

b. either he misses me or he doesn't love me

The association of a given attitude marker within a given factuality value is not entirely predictable. Sometimes, even the well-established identification of a certainty attitude with a factual value, which is posited as an axiom, for example in FactBank (Sauri and Pustejovsky, 2012), has to be reconsidered. Let us examine Example 4, where the non factual predicate "I think" and the certainty adverb "surely" impose respectively a non factual and a certainty value to the same event "there will come a time in my veterinary career that I don't get quite so ridiculous when confronted with a puppy"

(4) Sometimes I think that surely, eventually, there will come a time in my veterinary career that I don't get quite so ridiculous when confronted with a puppy.

Many annotation schemes tend to mix these two distinct notions. This is also the case in FactBank.

We claim that both from a theoretical point of view and because of the different purposes that an annotation of factuality and an annotation of speaker's attitude may have (factuality mining and opinion mining respectively) two different levels of annotation and two different annotation schemes should be provided for these two semantic dimensions. While this introduces a certain degree of redundancy, it also enhances clarity, flexibility, and completeness of the annotation, reflecting a theoretically valid distinction.

4 Annotation units

Factuality and speaker's attitude are often encoded by plenty of heterogeneous markers, both within a language and across languages (see also Morante and Sporleder (2012)). We believe that language-specific units of analysis should be determined only *after* cross-linguistic, functional categories have been defined. The lack of a functional background may lead to incomplete annotation schemes, if they are mainly based on the preliminary recognition of a set of markers prototypically connected with modality (such as modal verbs, modal adverbs or modal tags such as 'I guess'/'I believe'). Indeed, the cross-linguistic view of modality shows that there are various encoding strategies that can be overlooked by adopting a purely "lexical" approach.

Concerning factuality, for example, the non factual status of an event is determined not only by its occurence in the scope of a negation or a non factive predicate, but possibly also by an alternative coordinative construction (Mauri (2008)), see (3b) above.

Concerning speaker's attitude, future forms may function as epistemic markers with non-futural temporal reference, as exemplified by the English Future will in (Ex.(5), Nuyts (2006)) and by similar structures in German and in other Romance languages:

(5) Someone's knocking at the door. That will be John.

Similarly, past forms may be used as non-factual (specifically, counterfactual) markers (Fleischman (1995)) not only when they are under the scope of conditional markers (6a) but also when they are used in independent clauses (6b):

(6) a. Se lo sapevo venivo (Colloquial Italian) 'If I knew, I would come'

b. Io ero il principe e tu la principessa (Colloquial Italian) '(Let's pretend) I'm[past] the prince and you're the princess'

Modal particles are another common means for expressing modality. Though easily identifiable in texts, modal particles such as German 'denn' or English 'so' (Ex. 7a and 7b, De Haan (2006)) are somewhat neglected as triggers in the available annotation schemes, and this may be in part due to the difficult classification of their semantic contribution to the textual chunk containing them:

(7) a. Kommt er denn (German) 'Will he really come?'/'Will he come after all?'

b. There is so a Santa Claus!

As for the scope of the modal trigger, we claim that a distinction has to be made between factuality and speaker's attitudes. Factuality is a property of an event: it perfectly makes sense to attribute a factual status to each eventuality, as in Factbank. Speaker's attitude, instead, may apply to more or less extended spans of texts, ranging from a single word (8a) to a sequence of sentences (8b) and even to different dialogic turns (8c).

(8) a. It's a simple and (hopefully) nice cross-platform email chess program.

b. Hopefully he gets another shot and he finds a way to use this failure to motivate him to take the next step, to prove that guys like me completely underestimated him.

c. A: E' stato in banca? (Italian) Did he go to the bank? B: credo (Yes, I) think (so)

Current annotation schemes tend to consider the sentence as the domain within which the effects of a marker signalling the speaker's attitude are visible. Instead, we propose therefore not to aprioristically determine the scope of a trigger but to leave the annotator to identify it.

5 Implementation

The annotation model we are currently developing for both factuality and speaker's attitude is modular, language independent, and data-driven. The specific schemes for the annotation of triggers and markables are described below.

5.1 Schemes

Tables 1 and 2 show the annotation schemes for the elements *markable* and *trigger* respectively. Markables are all of the linguistic objects marked for *factuality* and all those marked for (speaker's) *attitude*. Triggers are those linguistic expressions that determine the factuality and attitude readings of the markables. Working with a functional layer allows us to use the same categories across languages. Markables are selected directly by the annotators and marked with the pre-specified attitude and factuality attributes, while linguistic realisations of triggers are pre-specified in a language-dependent fashion. Cross-language annotations can thus always be compared at the functional level, even in languages which code modality through very different linguistic expressions.

Table 1: Annotation categories for the *markable*

ATTITUDE	no		
	yes	epistemic	committment evidential
		deontic	manipulative volition
		valutative	axiological appreciative apprehensional
FACTUALITY	factual non_factual counterfactual		

The modal values in Table 1 are organised in a hierarchical structure, thereby allowing for a more flexible application of the annotation. If the annotator is uncertain about, say, the manipulative or volitional value of a markable (it could be the case for certain optatives, for instance), he can simply tag it as a deontic. If he cannot decide about the deontic or epistemic nature of a markable (which is often the case with possibilities), he can simply tag the mark-

Table 2: Annotation categories for the *trigger*, with examples of linguistic expressions which can be used in Romance (e.g. epistemic future) and Germanic languages (e.g. modal particles).

MORPHOLOGICAL	epistemic future reportive conditional other marker	
LEXICAL	verb	modal verb (which one) event selecting predicate (ESPs)
	noun	
	adjective	
	pragmatic_marker	adverbial parenthetical modal_particle connective question_tag
SYNTACTICAL	hypothetical alternative deontic	
OTHER		

able as a modalized linguistic object. We are confident that more fine-grained and coherent annotation can be driven from the annotation of real data, which should be regarded as an incremental dynamic task.

The left-hand column of Table 2 specifies categories that hold cross-linguistically. The linguistic realisations of triggers in the right-hand column are just examples which hold for some languages but would not (necessarily) be the same when considering other languages. Indeed, the annotation of triggers allows for both a general annotation of the syntactic nature of the trigger used (whether it is morphological, lexical or constructional in nature) and for a more language-specific annotation of the specific trigger used in a given language. Working this way has at least two advantages. First, we can compare different means of expressing same modality across languages. Second, we open the possibility of finding *prototypical*, or unmarked, linguistic expressions which serve as triggers for given modalities, much in the spirit of Croft (1991, 2000). Moreover, we think that such an approach may lead to interesting results for the automatic translation of modality.

5.2 Procedure and example

In the first stages of our annotation, we adopted the following procedure:

1. Identification of markables. We worked under the following assumptions:
   - these objects can vary for semantic nature and syntactic extension;
   - the linguistic objects marked for modality and those marked for factuality do not need coincide
2. Identification of triggers.
3. For each markable we specify:
   - its factuality value
   - its attitude value
   - the factuality trigger
   - the attitude trigger
4. For each trigger we specify:
   - its syntactic nature: a morphological element, a lexical element or a syntactic construction
   - the language-specific category used as a marker (for example the epistemic future for Romance languages, the mirative affix in Turkish, etc.)

As for the scope of markables, it should be clear that markables are often nested within each other: by avoiding a predetermination of the extension and

the nature of the markables, we can provide an annotation for each relevant element of our corpus, ranging from the entire text, to an embedded single word. Each markable is linked to its own trigger, regardless of the level of embedding of the trigger itself. Technically, this is done via layers of standoff annotation for factuality and attitude, which point to markables and triggers via their id value.

We use Example 9, from the Europarl corpus (Koehn, 2005), to illustrate our annotation procedure and schemes:

(9) In this respect, we should heed the words of von Eieck, and doubtless also those of the great Italian liberal Bruno Leoni, who warned precisely against the risks of an abnormal increase in anti-competition policies.

In Example 9 we can identify six markables:

(m1) we should heed [the words of von Eieck and doubtless also those of the great Italian liberal Bruno Leoni]

(m2) and doubtless also those of the great Italian liberal Bruno Leoni

(m3) who warned precisely against the risks of [an abnormal increase in anti-competition policies]

(m4) the risks of [an abnormal increase in anti-competition policies]

(m5) an abnormal increase [in anti-competition policies]

(m6) increase [in anti-competition policies]

They are marked up in text and then annotated for factuality and attitude according to the schemes described above in a standoff manner. For the sake of presentation, we show the annotation of markables and triggers separately in Figure 1, and the standoff annotation of attitude and factuality in Figure 2.

6 Conclusion and outlook

In our model we provide two independent annotation schemes for factuality and speaker's attitude, thus allowing for higher modularity and flexibility.

One of the main features of our model is the treatment of language specific markers of attitude and factuality as attributes of the modality type (which is instead language independent) assigned to each markable. This representation allows us on the one hand to separate the functional and the formal information, and on the other hand to specify how these are related to each other. This makes the proposed annotation scheme a powerful tool for investigating linguistic diversity in the field of modality on the basis of real language data, being thus also useful from the perspective of machine translation systems.

By avoiding a predetermination of the extension of markables and triggers, we can both provide an annotation for each relevant element of our corpus and account for the complex geometry of markables and triggers, which are often nested within each other. We believe that such an approach should improve the calculus of the percolation of modality along dependency trees and discourse relation structures.

The annotation schemes are being tested through manual annotation performed by expert annotators using existing tools such as GATE (Cunningham et al., 2011), MMAX (Müller and Strube, 2006), and BRAT (Stenetorp et al., 2012). Through annotation exercises and customisation we are currently exploring which might best suit our purposes. Intermediate evaluation of inter-annotator agreement is useful to identify inconsistencies in the scheme, and only after this first phase, the annotation will proceed on a larger scale. We are also considering existing collaborative platforms to perform distributed annotation over the web, so as to optimise the contribution of native speakers.

Content-wise, we plan to enrich our model in at least two ways: (1) by providing a coherent model for the annotation of the strength of modality values (certain, probable, impossible; necessary, prohibited, impossible, etc.); (2) by specifying for each modal attitude, the source of the attitude. Interannotator agreement will also be calculated to assess the validity of the scheme.

Concerning data, we are currently using the Europarl's parallel corpus (Koehn (2005)), but we also aim at including other comparable corpora to maximise linguistic diversity (languages outside Europe will be included) and register variation (mainly through the inclusion of spoken corpora).

```
In this respect , <markable id="m1">we should heed the words of von
Eieck, <markable id="m2">and doubtless also those of the great Italian
liberal Bruno Leoni</markable></markable> , <markable id="m3">who
warned precisely against <markable id="m4">the risks of <markable
id="m5">an abnormal <markable id="m6">increase in anti-competition
policies </markable></markable></markable></markable> .
```

```
In this respect , we <trigger id="t1" type="lexical" subtype="verb"
expr="modal_verb"> should</trigger> heed the words of von Eieck, and
<trigger id="t2" type="lexical" subtype="pragmatic_marker" expr="adverb">
doubtless</trigger> also those of the great Italian liberal Bruno
Leoni , <trigger id="t3" type="syntactical" subtype="relative_clause"
expr="who+V">who <trigger id="t4" type="lexical" subtype="verb"
expr="event_selecting_predicate">warned</trigger> precisely against the
<trigger id="t5" type="lexical" subtype="noun">risks</trigger></trigger>
of an <trigger id="t6" type="lexical" subtype="adjective">abnormal
</trigger> increase in anti-competition policies .
```

Figure 1: Markable and trigger annotation of Example 9.

```
<annotation name="factuality">
<factuality ref="m1" value="nonfactual" trigger="t1"/>
<factuality ref="m3" value="factual" trigger="t3"/>
<factuality ref="m4" value="factual" trigger="t4"/>
<factuality ref="m5" value="nonfactual" trigger="t5"/>
</annotation>

<annotation name="attitude">
<attitude ref="m1" value="deontic" type="manipulative" trigger="t1"/>
<attitude ref="m2" value="epistemic" type="commitment" trigger="t2"/>
<attitude ref="m4" value="deontic" type="manipulative" trigger="t4"/>
<attitude ref="m6" value="valutative" type="apprehensional" trigger="t6"/>
</annotation>
```

Figure 2: Factuality and Attitude annotation for markables of Example 9. Values for pointers are those shown in the annotation in Figure 1.

References

Baker, K. et al. (2010). SIMT SCALE 2009 - Modality Annotation Guidelines, Technical Report. Johns Hopkins, Baltimore.

Baker, K., B. Dorr, M. Bloodgood, C. Callison-Burch, N. Filardo, C. Piatko, L. Levin, and S. Miller (2012). Use of modality and negation in semantically-informed syntactic MT. *Computational Linguistics 38*.

Croft, W. (1991). *Syntactic Categories and Grammatical Relations: The Cognitive Organization of Information*. University of Chicago Press.

Croft, W. (2000). Parts of speech as language universals and as language particular categories. In P. Vogel and B. B. Comrie (Eds.), *Approaches to the Typology of Word Classes*, pp. 65–102. Berlin/New York: Mouton de Gruyter.

H. Cunningham et al. 2011. *Text Processing with GATE (Version 6)*.

De Haan, F. (2006). Typological approaches to modality. In W. Frawley (Ed.), *The expression of modality*, pp. 27–69. Mouton de Gruyter.

Fleischman, S. (1995). Imperfective and irrealis. In J. L. Bybee and S. Fleischman (Eds.), *Modality in discourse and grammar*, pp. 519–551. John Benjamins.

Hendrickx, I., A. Mendes, and S. Mencarelli (2012). Modality in text: a proposal for corpus annotation. In *Proc. of LREC'12*.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, Phuket, Thailand, pp. 79–86. AAMT.

Masini F. and Pietrandrea P. (2010). Magari. *Cognitive Linguistics 21*(1).

Mauri, C. (2008). The irreality of alternatives. *Studies in Language*.

McShane, M., S. Nirenburg, and R. Zacharski (2004). Mood and modality: out of theory and into the fray. *Nat. Lang. Eng. 10*(1), 57–89.

Morante, R. and C. Sporleder (2012). Modality and negation: An introduction to the special issue. *Computational Linguistics 38*(2).

Christoph Müller and Michael Strube. 2006. Multi-level annotation of linguistic data with MMAX2. In S. Braun, K. Kohn, and J. Mukherjee, eds, *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*, pages 197–214. Peter Lang, Frankfurt a.M., Germany.

Nirenburg, S. and M. McShane (2008). Annotating modality. Tech. report, University of Maryland.

Nuyts, J. (2006). Modality: Overview and linguistic issues. In W. Frawley (Ed.), *The expression of modality*, pp. 1–26. Mouton de Gruyter.

Sauri, R. and J. Pustejovsky (2009). Factbank: a corpus annotated with event factuality. *Language Resources and Evaluation 43*(3), 227–268.

Roser Sauri and James Pustejovsky. 2012. Are you sure that this happened? Assessing the factuality degree of events in text. *Computational Linguistics*, 38(2):261–299.

P. Stenetorp, S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, and J. Tsujii. 2012. BRAT: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at EACL'12*, 102–107, Avignon, France.

Szarvas, G., V. Vincze, R. Farkas, and J. Csirik (2008). The bioscope corpus: annotation for negation, uncertainty and their scope in biomedical texts. In *Proc of BioNLP '08*, Stroudsburg, PA, USA, pp. 38–45.

Vincze, V., G. Szarvas, G. Móra, T. Ohta, and R. Farkas (2010). Linguistic scope-based and biological event-based speculation and negation annotations in the genia event and bioscope corpora. In N. Collier et al. (Eds.), *Proc of the Fourth Int. Symp. for Semantic Mining in Biomedicine*, Cambridge, UK.

Wiebe, J., T. Wilson, and C. Cardie (2005). Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation 39*, 165–210.