MODAL: A multilingual corpus annotated for modality

Malvina Nissim
CLCG, University of Groningen
The Netherlands
m.nissim@rug.nl

Paola Pietrandrea
University of Tours, CNRS UMR7270
France
pietrandrea-guerrini@univ-tours.fr

Abstract

English. We have produced a corpus annotated for modality which amounts to approximately 20,000 words in English, French, and Italian. The annotation scheme is based on the notion of *epistemic construction* and virtually language-independent. The annotation is rigorously evaluated by means of a newly developed strategy based on the alignment of the entire epistemic constructions as identified and marked up two annotators. The corpus and the agreement scoring tools are publicly available.

Italiano. *Presentiamo un corpus multilingue di circa 20,000 parole annotato per modalità epistemica. La procedura di annotazione è guidata dal concetto di costruzione epistemic. La validità dell'annotazione è valutata attraverso una strategia sviluppata per tenere conto della necessità di allineare intere costruzioni identificate da annotatori diversi. Il corpus e gli strumenti per la valutazione dell'annotazione sono resi disponibili.*

1 Introduction and Background

Modality is a pervasive phenomenon crucial to language understanding, analysis, and automatic processing (Morante and Sporleder, 2012). The creation of modality-annotated data would benefit Natural Language Processing in at least two major aspects: (i) factuality detection, consisting in the automatic distinction between propositions that represent factual events and propositions that represent non factual ones; and (ii) sentiment analysis, which involve the processing of extra-propositional aspects of meaning and the detection of polarised judgements. Additionally, the annotation of modality may also have important repercussions in the field of corpus linguistics, as the techniques developed in the automatic treatment of modality can be used to improve our linguistic knowledge of modality itself.

As far as the detection of polarised judgments goes, there have been substantial annotation efforts in recent years, exemplified by recurring and increasing sentiment analysis tasks within the context of the Semeval evaluation campaign.[1] Attention has also been given to more specific factuality tasks such as the CoNLL-2010 Shared Task on identifying hedges (Farkas et al., 2010), and factuality annotation in languages other than English, such as Italian (Minard et al., 2014), and Dutch (Schoen et al., 2014). However, these are annotation efforts involving specific phenomena rather than modality in general.

Indeed, a major bottleneck in the creation of modality-annotated resources is the very notion of modality itself, as encapsulating this phenomenon in one exhaustive but workable definition is far from trivial (Morante and Sporleder, 2012). Building on the function-based proposal advanced in (Nissim et al., 2013) and (Ghia et al., 2016), we have created a comprehensive annotation scheme for epistemic modality and have applied it to multiple languages. Contextually, we have developed and deployed an evaluation strategy which shows that the corpus is annotated reliably.

Summary of contributions We produced the first multilingual corpus annotated for modality. The annotation scheme is virtually language-independent, and the annotation is evaluated according to a specifically designed methodology

[1] http://alt.qcri.org/semeval2017/index.php?id=tasks. Note that in 2017 within the sentiment analysis track there was also a task on truth detection, which goes to show how closely related the two phenomena indeed are.

which is portable to other tasks where annotators are left with substantial freedom in the selection of the tokens to be marked up. The corpus and the tools for scoring agreement are publicly available (http://modal.msh-vdl.fr/,https://bitbucket.org/lennyklb/modality/).

2 Corpus

The MODAL Corpus is the first corpus of dialogues in multiple languages annotated for phenomena of (epistemic) modality.

MODAL consists of three equivalent resources of English, French and Italian dialogues. These were drawn from the Santa Barbara Corpus of Spoken American English (Du Bois et al., 2000) for English, from the ESLO Corpus (Baude and Kanaan, 2014), plus the OTG Corpus and the Accueil UBS Corpus (Antoine et al., 2002) for French, and from the VoLip Corpus (Alfano et al., 2014) for Italian. All data is marked for epistemic modality and amounts to approximately 20.000 words per language for a total of 2824 epistemic constructions (833 for the English Corpus, 1271 for the French Corpus, 720 for the Italian Corpus).

2.1 Approach to annotation

In the construction of MODAL, we were guided by two main principles: *maximum expressivity*, and *cross-lingual validity*. We therefore took an approach to annotation that would simultaneously ensure both.

Specifically, we did not want to annotate a predetermined list of epistemic constructions and assign functions to them. Indeed, this would make the scheme very much language-dependent, as specific tokens/constructions would need to be identified for each language. Additionally, it would restrict the annotation to this pre-selection, which could not be exhaustive.

As an alternative approach, we provided a theoretical meaningful, and operationalisable definition of epistemic modality. On this ground, thus only at a later stage, the annotators identified the linguistic constructions that realise epistemic modality in the three different languages. Thus, rather than going from constructions to functions, we go from functions to constructions.

While this approach has the advantage of being valid cross-linguistically, and maximising expressivity, it also potentially has a major problem. Letting the annotators choose freely the tokens and the constructions to be annotated without controlling for any pre-selection, incurs the risk of a wide range of choices, and substantially low agreement. We discuss this in the Evaluation section. In the remainder of this section we explain the scheme and the procedure we used to annotate the corpus.

Table 1: Annotation categories for the marker

LEMMA	$< lemma >$
ILLOCUTION	assertion exclamation injunction question
MORPHOSYNTAX	morph-conditional morph-preterite morph-future lex-complement-taking-pred lex-adverb lex-disc-marker lex-modal-verb syn-dependent syn-list syn-tag disc-utterance prosody-interrogative

Table 2: Annotation categories for the Relation

DIRECTION	scope-marker marker-scope inside co-extensive	
EPISTEMIC TYPE	direct-auditory direct-visual direct-feeling indirect-infer indirect-report quotative memory no evidence	
POLARITY	positive neutral negative	
DISCOURSE FUNCTION	qualification negotiation	 acceptation non acceptation check information

2.2 Procedure and final scheme

We employed a two-fold procedure: epistemic constructions are first identified, and then annotated with their features.

Identification of epistemic constructions In order to annotate epistemic modality in *dialogues*, we subscribed to a communitarian (Stalnaker, 1978), dynamic (Groenendijk and Stokhof, 1991), and interactionist (Ginzburg, 2012) approach to semantics, which led us to refine the traditional definition of epistemic modality. Specifically, we put forward the idea that any construction that explicitly signals the process of shared attribution of a truth value to the propositional tokens that compose a discourse should be considered as an epistemic construction, and thus annotated.

Consequently, we annotated not only constructions in which a marker is realized by a more grammaticalized element, such as a modal verb (Example 1), but also constructions in which a marker is realized lexically (Example 2) or prosodically (Example 3):

(1) A penguin might lay two eggs and at that point [...]

(2) And I do believe it was thirty days [...]

(3) DON: Oh specifically in the islands?

Besides, we annotated both monological epistemic constructions in which a marker expresses the evaluation of the truth-value of a scope by a single speaker (Example 4), and dialogical epistemic constructions in which two or more markers are used to negotiate the evaluation of the truth-value of one and the same scope among the participants in a conversation (Example 5):

(4) apparently it was very very muddy it was abnormally warm and it was just a big mudbath out there [...]

(5) ALIC: I don't think Darren put anything on it .
NICO: Mhm .
ALIC: Right .
ALIC: Okay .

Annotation of epistemic constructions We represented the epistemic constructions identified in the corpus as triadic constructions consisting of a marker, a scope and a relation between the marker and the scope, as shown in Figure 1.

We formalised the marker, the scope and the relation between them as three elements each endowed with its own formal and functional properties. Each element is then annotated with syntactic, semantic, and pragmatic features according to the developed annotation scheme.

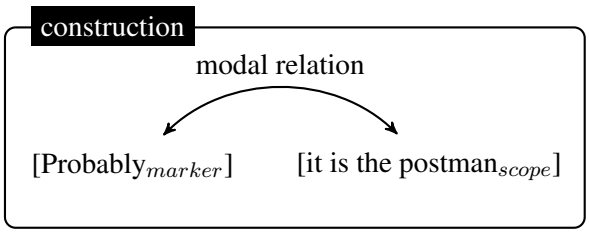

Figure 1: A construction is conceived as a marker, a scope, and a modal relation between them.

Building on (Nissim et al., 2013; Ghia et al., 2016), we devise a fully-fledged annotation scheme that is functionally motivated and cross-linguistically valid. Annotation features are specified for all three elements of the modalised construction, namely the marker, the scope, and the relation. The features for the markers and the relations are shown in Tables 1 and 2, respectively. For the scope, we use a property `syntax` and `clause`/`utterance` as features.

Operationalising the annotation task From a theoretical perspective, the scheme is grounded in the Construction Grammar framework (Goldberg, 1995). In practice, the annotators could work with the labels from the annotation schemes, but also with decision trees that guided the process of identification of epistemic constructions as well as feature assignment. The annotation was performed using the Analec annotation tool (Landragin et al., 2012), which produces TEI-compliant XML output. Analec was originally designed for the annotation of anaphoric phenomena and thus lends itself well to the task of annotating a three-way construction, with features for marker, scope, and relation. All data was annotated by three teams of 2 or more annotators (a, b for Italian, a, b, c, d for English, a, e,f for French) and agreement was assessed via a specifically developed evaluation strategy (Section 3).[2]

3 Evaluation

The originality of the general approach and of the annotation procedure led us to develop an origi-

[2]Further information regarding the distribution of categories and examples is available at the project's website (`http://modal.msh-vdl.fr/` and in (Pietrandrea, forthcoming)).

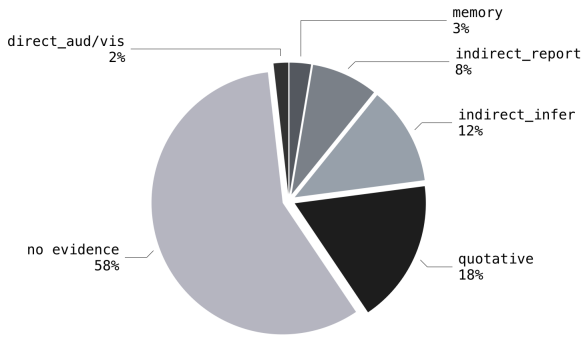

Figure 2: Distribution of EPISTEMIC TYPES for the Relation annotation in the Italian portion of the corpus (see Table 2).

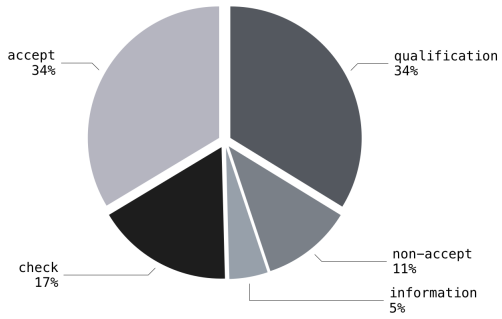

Figure 3: Distribution of DISCOURSE FUNCTIONS for the Relation annotation in the Italian portion of the corpus (see Table 2).

nal technique for testing the inter annotator agreement, essentially based on the percentage of overlap between the spans of text identified as markers or scopes by the annotators (Ghia et al., 2016). In order to assess this, annotations must be *aligned*. We describe how we align the constructions in practice, how we use alignment information in order to assess agreement, and discuss results.

3.1 Alignment and Agreement

Annotators can identify *any* textual element as part of a modalised construction, and each annotator works on their own file. This means that in order to assess agreement, we first need to try to *align* the constructions marked up in the two files. We do so via *anchors*. Anchors can be aligned iff:

- they are of the same type (marker or scope)
- they overlap in content by at least a given proportion of lexical material, which we base on character offset. For example, for a required overlap of 50% and a token length of an anchor A of ten tokens, the content of the candidate anchor from the other file needs to have at least five subsequent words in common with A.

This process results in a collection of pairs of aligned anchors. For example, considering annotator a and annotator b, we would have an aligned pair of marker t_a and marker t_b.

The final step is to iterate through the relations that judge a introduced and align them with relations that judge b introduced. In order to explain the procedure of further alignment to relations, we take judge a as reference, but in terms of scores it doesn't make any difference which direction we go, since $precision_{ab} = recall_{ba}$ so that eventually $fscore_{ab} = fscore_{ba}$. Relations consist of a marker and one or multiple scope portions. Aligning relations is done by pairing up markers and scopes into relations introduced by judge a and check if the aligned counterparts of these markers and scopes by judge b are part of a relation as well. When this is the case, we deem the two constructions as "the same".

Next, we have to assess agreement on the features assigned to relations and markers. While agreement over alignment is measured using precision/recall/f-score as we have to deal with potentially different spans, for the relations' and markers' features, we can then use Cohen's Kappa (Cohen, 1960) over the agreed upon constructions only, as it becomes a plain classification task.

3.2 Results

Because of freedom in the annotation of the extension of anchors, as mentioned above we evaluated alignment at different percentages of overlap. The scores for the alignment of scopes for all three languages is shown in Table 3. While for Italian we observe that even when evaluating alignment of full strings (i.e. requiring 100% overlap), the agreement stays high, this is not the case for French and English. Indeed, if complete overlap of scopes is required to deem the annotations equivalent, F-scores drop quite a bit. We do not include a table for the scores on the markers as they do not change substantially with varying degrees of overlap. This is due to the fact that markers are often just single words, or very short anyway. F-scores range from 0.91 at 10% and 0.90 at 100% for English, from 0.86 at 10% and 0.85 at 100% for French, and stay stable at 0.94 for Italian. For this reason, we can be lenient with markers' align-

Table 3: Agreement for scope identification.

Overlap	FRENCH			ITALIAN			ENGLISH		
	Prec	Rec	F	Prec	Rec	F	Prec	Rec	F
10%	0.86	0.92	0.89	0.95	0.93	0.94	0.88	0.85	0.86
20%	0.85	0.92	0.88	0.95	0.93	0.94	0.88	0.85	0.86
30%	0.84	0.92	0.88	0.95	0.93	0.94	0.88	0.85	0.86
40%	0.82	0.92	0.87	0.95	0.93	0.94	0.88	0.85	0.86
50%	0.82	0.92	0.87	0.95	0.93	0.94	0.88	0.85	0.86
60%	0.79	0.91	0.85	0.95	0.93	0.94	0.88	0.85	0.86
70%	0.78	0.9 0	0.84	0.95	0.93	0.94	0.87	0.84	0.85
80%	0.77	0.9 0	0.83	0.95	0.93	0.94	0.87	0.83	0.85
90%	0.75	0.89	0.81	0.94	0.93	0.93	0.85	0.81	0.83
100%	0.72	0.87	0.79	0.94	0.93	0.93	0.81	0.78	0.79

Table 4: Kappa scores for marker's features.

Overlap	FRENCH	ITALIAN	ENGLISH
10%	0.82	0.91	0.85
20%	0.82	0.91	0.86
30%	0.84	0.91	0.87
40%	0.84	0.91	0.88
50%	0.84	0.91	0.88
60%	0.84	0.91	0.88
70%	0.84	0.92	0.87
80%	0.84	0.92	0.87
90%	0.89	0.92	0.86
100%	0.89	0.92	0.87

ment, which we set at 10% when evaluating relations. Interestingly, though, we can observe that when evaluating agreement on the *features*, Kappa increases when stricter alignment is required (Table 4). This is likely due to the fact that on fully agreed upon strings, the assigned features are also agreed upon.

For the Italian annotation of the relation's features, at overlap 100%, we observe $K = 0.86$ for the FUNCTION feature, $K = 0.82$ for TYPE, and $K = 0.72$ for POLARITY.[3]

To provide a more detailed view into the disagreements of the `type` feature, for instance (whose final agreed upon distribution was reported in Figure 2 above), in Figure 4 we show the confusion matrix for Italian. We can observe that the largest number of confusions arise from mixing up the categories `indirect_inferential` and `no_evidence`. Indeed, the precise delimitation between these two categories is a long- and hot-debated issue in the literature on epistemic modality (Pietrandrea, 2005, among others).

Overall, we can see that our annotation, albeit granting the annotators a lot of freedom, is substantially reliable.[4]

[3]Very similar scores are observed at different degrees of overlap.

[4]Please note that for all agreement results, for all languages, the reader is referred to the project's website, where examples of disagreement are also included (`http://modal.msh-vdl.fr/`).

	dir_aud	dir_vis	ind_inf	ind_rep	mem	no_ev	quot
dir_aud	**1**	-	-	-	-	-	-
dir_vis	-	**9**	-	-	-	-	-
ind_inf	-	-	**40**	2	-	21	1
ind_rep	-	-	-	**46**	-	1	1
mem	-	-	2	-	**12**	1	-
no_ev	-	1	12	3	1	**316**	3
quot	-	-	-	7	-	7	**90**

Figure 4: Confusion matrix for the annotation of the TYPE feature in Italian.

4 Conclusions

Modality can be reliably annotated in multiple languages by taking a bottom-up, functional approach paired with a solid annotation scheme, trees to guide the annotators' decisions, and a rigorous evaluation strategy. With this approach, we have produced the first multilingual corpus annotated for modality, which can be potentially used to train modality detection models as well as to further study modality itself. By making all of the data publicly available, and by sharing our annotation experience, we also hope to provide a blueprint for creating modality-annotated resources in yet more languages.

Acknowledgments

The research described in this paper was supported by the Maison des Sciences de l'Homme du Val de Loire (grant MoDAL 2015), the IRCOM Consortium, the "Laboratoire Ligérien de Linguistique – UMR7270", and the Groningen Meaning Bank. The authors are also grateful to the anonymous reviewers for their comments.

References

Iolanda Alfano, Francesco Cutugno, Aurelio De Rosa, Claudio Iacobini, Renata Savy, and Miriam Voghera. 2014. Volip: a corpus of spoken italian and a virtuous example of reuse of linguistic resources. In Nicoletta Calzolari et al., editor, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

J.-Y. Antoine, S. Letellier-Zarshenas, P. Nicolas, and I. Schadle. 2002. Corpus OTG et ECOLE_MASSY : vers la constitution dun collection de corpus francophones de dialogue oral diffusés librement. In *Actes TALN2002*, pages 319–324, Nancy, France.

Olivier Baude and Layal Kanaan. 2014. Le corpus des ESLO à l'ère des Digital Humanities. In *Premières Rencontres FLORAL*, Paris, France, December.

J. Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37.

John W Du Bois, Wallace L Chafe, Charles Meyer, Sandra A Thompson, and Nii Martey. 2000. Santa barbara corpus of spoken american english. *CD-ROM. Philadelphia: Linguistic Data Consortium*.

Richárd Farkas, Veronika Vincze, György Móra, János Csirik, and György Szarvas. 2010. The CoNLL-2010 shared task: learning to detect hedges and their scope in natural language text. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning — Shared Task*, CoNLL '10: Shared Task, pages 1–12, Stroudsburg, PA, USA. Association for Computational Linguistics.

E. Ghia, L. Kloppenburg, M. Nissim, P. Pietrandrea, and V. Cervoni. 2016. A construction-centered approach to the annotation of modality. In *Proceedings of the 12th Workshop on Interoperable Semantic Annotation (ISA-12)*, Portoroz, Slovenia.

Jonathan Ginzburg. 2012. *The interactive stance*. Oxford University Press.

Adele E Goldberg. 1995. *Constructions: A construction grammar approach to argument structure*. University of Chicago Press.

Jeroen Groenendijk and Martin Stokhof. 1991. Dynamic predicate logic. *Linguistics and philosophy*, 14(1):39–100.

Frederic Landragin, Thierry Poibeau, and Bernard Victorri. 2012. Analec: a new tool for the dynamic annotation of textual data. In *International Conference on Language Resources and Evaluation (LREC 2012)*, pages 357–362.

Anne-Lyse Minard, Alessandro Marchetti, and Manuela Speranza. 2014. Event Factuality in Italian: Annotation of News Stories from the Ita-TimeBank. In *Proceedings of CLIC-it*, Pisa.

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

Malvina Nissim, 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, March. ACL.

Paola Pietrandrea. 2005. *Epistemic modality: functional properties and the Italian system*, volume 74. John Benjamins Publishing.

Paola Pietrandrea. forthcoming. Epistemic constructions at work: A corpus-driven study on spoken Italian dialogues. *Journal of Pragmatics*.

Anneleen Schoen, Chantal van Son, Marieke van Erp, and Hennie van der Vliet. 2014. Newsreader document-level annotation guidelines-dutch. Technical report, Vrije Universiteit Amsterdam, TechReport 2014-8.

Robert C Stalnaker. 1978. *Assertion*. Wiley Online Library.