A Construction-centered Approach to the Annotation of Modality

Elisa Ghia[1], **Lennart Kloppenburg**[2], **Malvina Nissim**[2], **Paola Pietrandrea**[3], **Valerio Cervoni**[3]
[1]University for Foreigners of Siena, [2]University of Groningen, [3]University of Tours and CNRS LLL
elisaghia@gmail.com, {l.kloppenburg|m.nissim}@rug.nl, pietrandrea-guerrini@univ-tours.fr, cervoni@etu.univ-tours.fr

Abstract

We propose a comprehensive annotation framework for modality, which encompasses and supports existing annotation schemes, by adopting a construction-centered view. Rather than seeing modality as a feature of a trigger or of a target, we view it as a feature of the triad "trigger-target-relation", which we name *construction*. We motivate the need for such an approach from a theoretical perspective, and we also show that a construction-centered annotation scheme is operationally valid. We evaluate inter-annotator agreement via a pilot study, and find that modalised constructions identified by different annotators can be successfully aligned, as a first crucial step towards further agreement evaluations.

Keywords: modality, annotation, agreement

1. Introduction

Modality is a pervasive phenomenon crucial to language understanding, analysis, and automatic processing, and at the same time difficult to encapsulate in one exhaustive but workable definition (Morante and Sporleder, 2012). This is reflected in the continuous efforts towards two intertwined aims, namely (i) the definition of the core and the borders of modality, and (ii) the creation of annotated data, also towards the development of automatic systems.

Indeed, 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) opinion mining and sentiment analysis, which involve the processing of extra-propositional aspects of meaning and the detection of polarised judgements. Efforts in this sense are exemplified by recurring sentiment analysis tasks within the context of Semeval (see for example Task 9 to Task 12 in the 2015 campaign)[1], as well as specific factuality tasks such as the CoNLL-2010 Shared Task on identifying hedges (Farkas et al., 2010), and data annotation towards further campaigns, not just limited to English (Minard et al., 2014; Schoen et al., 2014).

In addition to NLP applications, the annotation of modality may have important repercussions in the Corpus Linguistics field, as the techniques developed in the automatic treatment of modality can be used to improve our linguistic knowledge of modality itself. Nevertheless, shared standards for modality annotation do not exist as yet (Morante and Sporleder, 2012).

In the current contribution, we apply the model described in (Nissim et al., 2013) to epistemic modality, and we describe the development and implementation of a flexible and comprehensive scheme for the annotation of modalised constructions in transcribed dialogues. With a view to developing a flexible model for the automatic annotation of modality, we suggest that the annotation procedure follow a corpus-driven approach, as operational categories can be drawn and refined from data. Because such a model has to be not only theoretically sound but operational (both in terms of annotation as well as in terms of automatic processing), we propose a comprehensive annotation framework for modality which we motivate theoretically, and test its validity empirically by means of a pilot study.

[1] http://alt.qcri.org/semeval2015/index.php?id=tasks

2. Phenomena

Annotating modality may involve the identification of factuality and/or subjectivity. These two dimensions are key not only to interpreting modalised constructions but also in terms of their annotation. Indeed, to summarise what we explain in detail below, approaches that focus more on the factuality aspect of modality are target-centered, while approaches that focus more on subjectivity (including here opinion mining) are trigger-centered (see Figure 1 for an example of trigger and target).

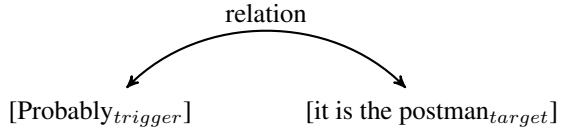

Figure 1: Trigger, target, and relation between them in a modalised context.

2.1. Factuality and Target-centered Schemes

Factuality refers to the extent to which the event described in a proposition is grounded in reality. Factuality annotation is hence aimed at distinguishing linguistic material presented as a fact from other language material. This has also to do with speculation (Medlock and Briscoe, 2007) and uncertainty (Rubin, 2010; Szarvas et al., 2012; Saurí and Pustejovsky, 2012; Sanchez and Vogel, 2015; Thompson et al., 2008). When focusing on factuality, the annotation is usually target-driven. This means that the annotation procedure consists in identifying the element whose factuality has to be evaluated, i.e., the target of the factuality relation, and in providing information about that element. In the annotation of FactBank (Saurí and Pustejovsky, 2012), for example, the text is segmented in 'events'. For each event the

schema specifies the following attributes: source, source introducing predicate, factuality value, time (see Figure 2). Building on the idea that the factuality of a semantic state can be annotated via the annotation of opinions, Wiebe et al. (2005) identify in the text the semantic entities that correspond to private states. Each private state is then annotated for intensity, attitude type, source, anchor text, target (Figure 3). In Thompson et al. (2008)'s annotation scheme the text is segmented in sentences. For each sentence the scheme specifies the certainty trigger that determines the factuality value of the sentence. For each trigger three attributes are specified: the point of view, the knowledge type and the certainty level (Figure 4).

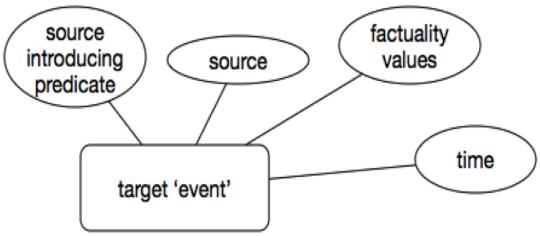

Figure 2: Overview of Saurí and Pustejovsky (2012)'s target-centered annotation scheme.

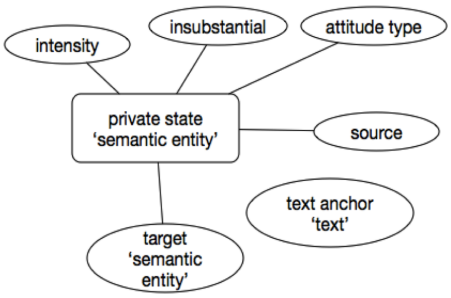

Figure 3: Overview of Wiebe et al. (2005)'s target-centered annotation scheme (in the context of opinion mining).

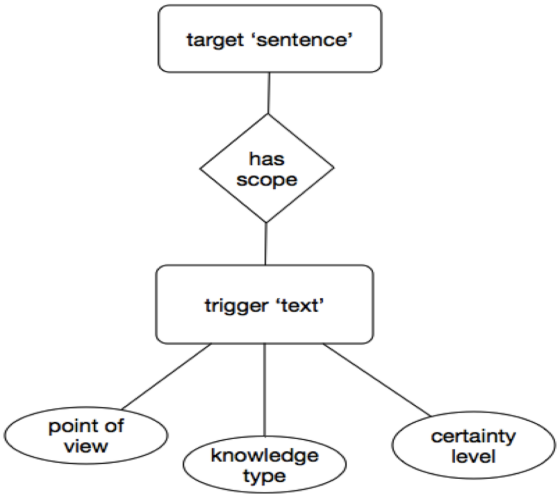

Figure 4: Overview of Thompson et al. (2008)'s target-centered annotation scheme.

2.2. Subjectivity and Trigger-centered Schemes

Beside factuality, modality interpretation involves the identification of subjectivity, or *extrapropositional aspects of meaning*. When specifically annotating subjectivity normally a wide notion is adopted, including such components as appreciation, fear, effort, epistemic opinion. Annotation schemes that focus on this aspect, including work on sentiment analysis, adopt a trigger-centered approach to annotation, as it is the subjectivity/sentiment of triggers that is mostly informative (Wiebe et al., 2005; Rubin, 2010; Nirenburg and McShane, 2008; Hendrickx et al., 2012; Ávila et al., 2015; Baker et al., 2010).

In brief, trigger-driven annotation approaches consist in identifying in a text the linguistic elements that encode the subjective meaning. Vincze et al. (2010)'s procedure, for example, consists in annotating the 'lexical cue' that encodes uncertainty and in specifying for each lexical cue the following attributes: the genre and the domain of the text in which it occurs, the type of uncertainty that it encodes, its PoS and the chunk it belongs to (Figure 5). Sanchez and Vogel (2015) take the 'hedges' encoding the degree of commitment of the speaker as the central element of their annotation. For each hedge they specify: the syntactic type, the dependency tree over which the hedge scopes, the type of source to which the hedge has to be attributed (Figure 6).

The schemes represented in Figures 2 to 6 specify different abstract syntaxes. However, from a conceptual standpoint all these schemes regard a modal relation in the same way, namely as a dyadic relation between a trigger and a target. According to the objectives of the annotation, either the trigger or the target of the relation is taken as the annotable unit.

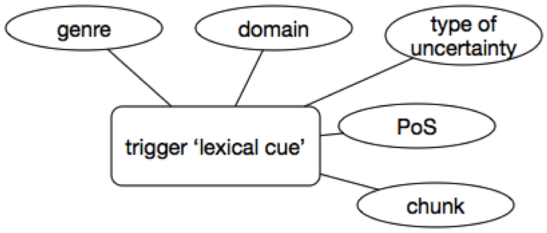

Figure 5: Overview of Vincze et al. (2010)'s trigger-centered annotation scheme.

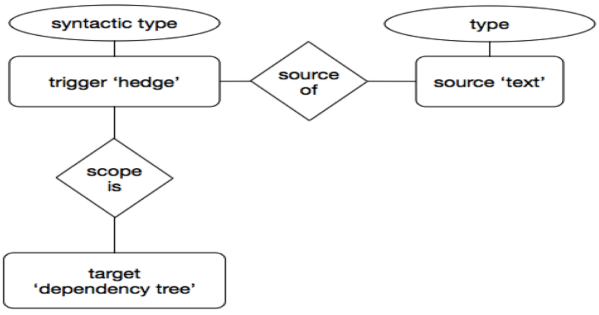

Figure 6: Overview of Sanchez and Vogel (2015)'s trigger-centered annotation scheme.

3. A Construction-centered Approach

Rather than as binary relations between a trigger and a target, in this contribution we view *constructions* as triadic relations between a trigger, a target and the relation between them. Accordingly, we revise Figure 1 as Figure 7.

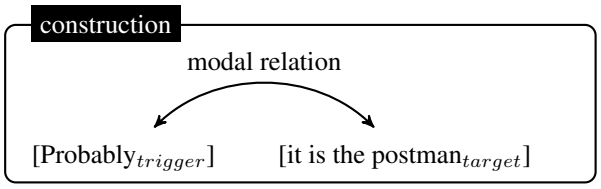

Figure 7: A construction is conceived as a trigger, a target, and a modal relation between them.

As a consequence, from a conceptual point of view, the relation between the trigger and the target has its own properties and functions, and from a practical perspective, such properties and functions have to be specified as attributes of the relation itself in an annotation scheme. In what follows, we justify this view theoretically (Section 3.1.) as well as practically (Section 4. and Section 5.).

3.1. Theory

Our formalisation choice is linked to multiple factors. First, it happens quite frequently in spoken language (and it may theoretically happen in written language alike) that one and the same target can receive more than one evaluation. In Example (1):

(1) C: dovevano venire a leggerla quanto meno [they were supposed to come and read it, at least]

A: no anche se non veniva<no> sì dovevano veni' a leggerla [no even if they didn't come- yes they were supposed to come and read it]

C: cosi' almeno si sapeva [at least this is what we knew]

the same target, i.e. the proposition [they come and read it] is linked to four different truth values through four different triggers:

1. the modal "dovevano" ['were supposed to']
2. the pragmatic marker "no" ['no']
3. the pragmatic marker "sì", ['yes']
4. the utterance "cosiÕ almeno si sapeva" ['at least this is what we knew'].

In other words, the target enters four different epistemic constructions, and it would not make sense to try to establish its factuality status value independently of such constructions. As a consequence, factuality evaluation cannot be conceived as a property of the target (which indeed receives several modality evaluations).

It would be equally awkward to regard modal evaluation as a property of the trigger. As the utterances in Examples (2) through (6)[2] show, one and the same trigger – in this case the complement-taking predicate "I think" – triggers different types of modality to its target, according to the target's semantic nature, whether a statement, a judgement, etc.

(2) I think he went through a separation with his wife and I think that depressed him.

(3) I love your wife, and I think she is beautiful!

(4) Quite frankly, I think he has the right to make that decision.

(5) I think you are better off fixing the "issues" one by one than going into bankruptcy.

(6) I did have a waxing service from one other person here, but I think I will choose to stick with Simona for future waxing services from here.

Therefore, modal evaluation is better regarded as a property of the construction as a whole, i.e. as a functional property encoding the relation between the trigger and the target. Our annotation scheme is grounded in such assumption.

3.2. Overview of the Annotation Scheme

We describe in this section the procedure and the scheme we adopted for our annotation task.

As a first step, we identified triggers and targets in a modalised construction in the text. Once selected, triggers were defined through the attributes form (i.e. text token), lemma, illocution (i.e. the trigger's illocution, involving the values assertion, expression, injunction, question) and morphosyntactic category (including morphological triggers, e.g. tense/aspect marking, lexical triggers, e.g. adverbs or pragmatic markers, syntactic triggers, such as inversions for interrogatives, prosodic triggers). The target was subsequently identified and defined by its illocution (assertion, exclamation, injunction, question). The third stage in the annotation procedure involved the linking of trigger and target into a modal relation, which was further defined through the attributes direction (trigger $>$ target or target $>$ trigger, embedding, co-extension, extension over more turns and speakers), function (i.e. discourse function: qualifying, accepting, non accepting, checking, confirming, non confirming, informing), polarity (positive, negative, neutral), and type (type of evidence upon which the epistemic construction is grounded). See (Pietrandrea, submitted) for a theoretical justification of the annotation schema (see Figure 8 for a screenshot of the annotation schemes with all of the categories and labels).

As we have seen, different approaches are associated with different schemes which respond to different objectives. However, even distinct modal phenomena are related to each other, as they all deal with the validation of representations or, in other words, with extrapropositional aspects of meaning. A flexible and broader annotation scheme could thus allow to encompass all specific schemes and support the needs of target-, trigger-, and relation-centred approaches altogether.

The comprehensive scheme was tested on the annotation of modalised constructions in spoken text. Annotation was carried out by multiple annotators on the basis of guidelines

[2]These examples are all from the EnTenTen corpus (Pomikálek et al., 2009).

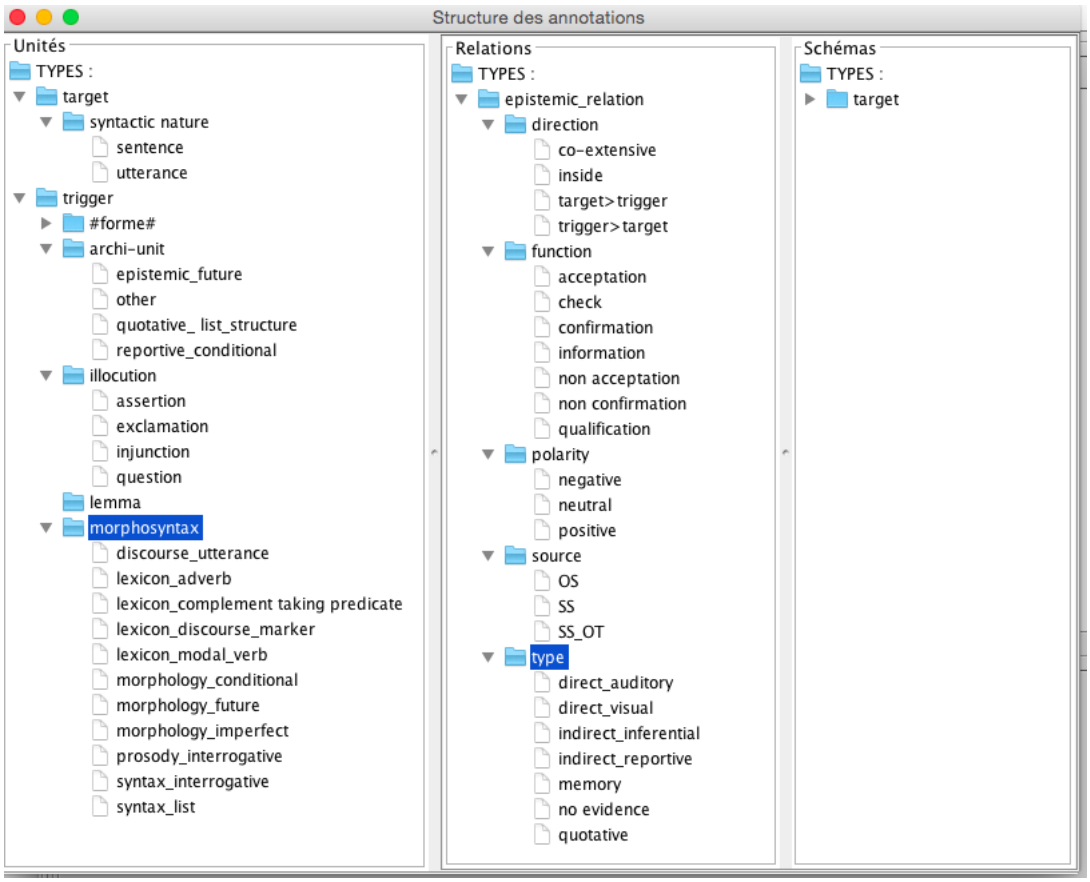

Figure 8: Screenshot of the Analec annotation tool customised for a construction-centered modality annotation. The categories and labels of the annotation scheme are all visible.

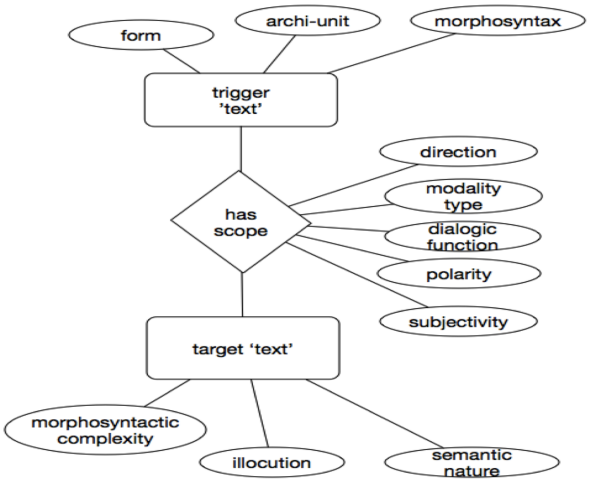

Figure 9: Overview of our construction-centred annotation scheme.

established via decision trees. A common and highly customizable annotation tool was used for manual annotation, along with shared evaluation metrics. Annotators discussed the annotation process and the operational categories at regular meetings.

After initial identification of the relevant categories by multiple annotators, the cognitive salience of such categories is recursively tested through inter-annotator agreement. The model is hence refined incrementally (Glynn and Krawczak, 2014), leading to ultimate operationalisation of the categories that allow for the semantic modelling of modality.

4. Annotation Experiment

With a view to testing our annotation framework for modality and its implementation on language data, annotation has proceeded along a set of successive stages: (i) (pilot) annotation by multiple expert annotators and identification of relevant categories, (ii) calculation of inter-annotator agreement to test the feasibility of a construction-based annotation (tested via alignment, see below) and the cognitive salience of the categories, (iii) refinement of the annotation scheme, (iv) second annotation phase, (v) operationalisation of the necessary categories for the semantic modelling of modality. We are describing here stages (i) and partially (ii)-(iii).

The pilot experiment is divided into two phases, and involves the annotation of spoken Italian data from the LIP corpus (De Mauro, 1993) and spoken French from the ESLO corpus[3]. The annotation was performed using the Analec annotation tool (Landragin et al., 2012), which produces TEI-compliant XML output, and was originally designed for the annotation of anaphoric phenomena and thus lends itself well to the task of annotating a three-way

[3] http://eslo.huma-num.fr/

construction, with features for trigger, target, and relation. Appropriate categories and features were implemented via in-tool customisation of the annotation schemes (see Figure 8).
Each annotator worked individually following these steps:

1. identification of the trigger of the modal construction
2. identification of the target of the modal construction
3. identification of the relation holding between trigger and target.

For the first phase, on Italian, a total of approximately 650 constructions was annotated on a corpus section of 19,665 words consisting of six dialogic situations: a university exam, a dialogue excerpt from a television programme, a transactional exchange, two conversations among friends, one family conversation over dinner. At this stage, the alignment of constructions across annotations, needed to assess whether the judges had identified the same modalised constructions, was performed in a rather simple and shallow way, with substantial manual intervention. Note that alignment is crucial towards assessing the validity of the scheme, as freedom must not imply randomness or the impossibility to perform evaluation. Although the annotation yielded promising results on agreement, with an f-score of 0.779 over the constructions, and 53% agreement on the exact extension of the targets, the alignment procedure wasn't properly formalised in any way. For a second pilot study we therefore refined the annotation guidelines not only conceptually but also operationally in order to provide more precise instructions regarding the spans to be annotated, and we devised a more structured, more robust and at the same time more flexible procedure for aligning annotations. This procedure is described in the next section, and has been deployed on portions annotated in the second phase of the pilot study.
This second phase focuses on French (also with a view to keep the scheme cross-linguistically valid), for which we are annotating 20,000 words in dialogues from the ESLO corpus. For this pilot experiment, 7 annotators working on a 1000 word text, annotated about 40 constructions.[4] For this showcase evaluation we take the annotations performed by two judges, a and b in what follows.

5. Evaluation

Before comparing and evaluating the values attributed to the relations and the triggers, we need to *align* the *constructions* identified by two annotators. Thus, in order to compute agreement among two annotated documents, the output XML files are subjected to preprocessing, alignment and agreement phases. The alignment phase is particularly meaningful, as it tells us whether two judges have identified the same moralised construction (independently of the specific feature values assigned to trigger/target/relation).

[4]Some coding details have changed, and we haven't yet transferred the whole Italian annotation to the current format, so that the more structured evaluation of alignment hasn't been performed on this data again, yet.

5.1. Preprocessing

The file is pre-divided into different paragraphs. Every paragraph contains either no or one/multiple trigger/target annotations, which we term *anchors* in this context. As a first step, we collect all of the anchors and extract the transcript contents between the beginning and end of an anchor, in other words: the marked up text. For example, in Figure 10, for **'u-trigger-3'**, the content (text) is *'il m'a dit'*, and for **'u-target_portion-3'** it is *'il travaillait pas'*. These anchor-content pairs are subsequently stored for both annotation files.

5.2. Alignment

The IDs of the anchors (displayed as 'id' in the XML-sample in Figure 10) of either annotation file do not correspond to each other as they obviously only obey internal consistency, and the texts are annotated separately. Therefore, we need an alignment step which matches anchors from both annotation files. Anchors can be aligned iff:

- they are of the same type (trigger or target)
- 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 (see Section 5.4. for an example of partial overlap and a further discussion of varying overlap requirements)

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 trigger t_a and trigger 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 trigger and one or multiple target portions. Aligning relations is done by pairing up triggers and targets into relations introduced by judge a and check if the aligned counterparts of these triggers and targets by judge b are part of a relation as well. In case this is the case, we deem the two constructions as "the same". Note that at this stage we have not checked yet agreement on the features assigned to relations and triggers – we are just evaluating that the two judges identify in text the same modalised construction, which is a crucial step.
The alignment process between judge a and judge b results then in three sets that can be evaluated: a set of *trigger* pairs, a set of *target* pairs and a set of *relation* pairs. The agreement between judge a and judge b for a given set is expressed as the precision of annotations by judge b compared to those of judge a. Recall for this same process is computed by swapping judge b and judge a, since as hinted above, a false negative, or a relation/trigger/target which was annotated by judge a but not by judge b, turns into a false positive if this is reversed.

```
<anchor id="u-trigger-3-start" type="AnalecDelimiter" subtype="UnitStart"/>
il m'a dit
<anchor xml:id="u-trigger-3-end" type="AnalecDelimiter" subtype="UnitEnd"/>
<anchor xml:id="u-target_portion-3-start" type="AnalecDelimiter" subtype="UnitStart"/>
il travaillait pas
<anchor xml:id="u-target_portion-3-end" type="AnalecDelimiter" subtype="UnitEnd"/>
```

Figure 10: Example annotation

5.3. Layers of Inconsistency

In the alignment process, there are a number of layers where mismatches and actual disagreements can occur. The *paragraph layer* refers to the possibility that any paragraph annotated by judge a is not annotated by judge b, which means that all annotations that judge a made in this particular paragraph cannot be aligned. This was necessary since in the pilot study not all annotators completed the whole text markup, thus leaving some final portions simply unannotated. Since this does not tell us anything about conceptual agreement, only the paragraphs which were annotated by both judge a and b were considered. This stage would not be relevant if complete texts are annotated by both annotators. At the *alignment layer* we align anchors. The process can fail if there is not enough overlap or if judge a annotated fewer or more anchors than judge b, which automatically results in failed alignments. The final layer is the *relation layer*. Consider that the alignment of two relations between judge a and judge b must obey the following constraints:

1. both the target and trigger of the relation by judge a need to be *aligned* with counterparts from judge b. If one of these was not aligned, the relation alignment fails as well

2. both of the counterparts need to belong to the same relation by judge b.

5.4. Results

According to the specific procedure just described, we report agreement results for construction alignment on a sample of two files from the French data, annotated by judge a and judge b, with approximately 40 constructions found.

As mentioned, we have to evaluate two main aspects. First, whether the annotators have identified the same modalised constructions, thus whether we can align their annotations. Second, whether the features assigned to triggers and to relations according to the annotation scheme correspond between the two judges. Agreement over alignment is measured using precision/recall/f-score as we have to deal with potentially different spans. For the relations' and triggers' features we can then use Cohen's Kappa (Cohen, 1960) (or Fleiss' Kappa in case of more than two annotators) over the agreed upon constructions only, as it becomes a plain classification task. In this paper, we are only reporting alignment agreement.

Because of freedom in the annotation of the extension of anchors, as mentioned above we evaluated alignment at different percentages of overlap. This is particularly relevant for the target portion. As for triggers, we can be very lenient, especially if the properties assigned to them by both annotators correspond.[5] For the pilot study that we present here, we have tested targets' overlaps in the range of 10% to 100% in terms of tokens.

To illustrate this, consider Example (7). The token overlap between the strings selected by the annotator a and annotator b is just under 50%. So setting the overlap constraint at 40% would yield an alignment and thus agreement, agreement, while setting at 50% wouldn't.

(7) a = 'il ne travaillait pas'
b = 'il m'a dit qu'il ne travaillait pas'

For the triggers, we have fixed the overlap at a minimum of 10%. At this level of overlap, f-score is measured at 0.87 (see Table 1), with a total of 34 aligned triggers out of 40 detected by a and 38 detected by b. By fixing this alignment for triggers, in Table 2 we report precision, recall, and the specific amount of true positives (TPs) and false negatives (FNs) at varying degrees of overlap. Results for alignment over constructions as wholes is given in Table 3.

Overlap	Prec	Rec	F1
10%	0.89	0.85	0.87
50%	0.84	0.85	0.84
100%	0.71	0.71	0.70

Table 1: Alignment agreement for *triggers* with varying amounts of overlap. Judge a is taken as reference in indicating precision and recall.

6. Conclusion

We have presented a construction-based annotation scheme for modality that is theoretically sound and empirically applicable. There are existing schemes that cover some aspects of modality annotation, but no specific shared standards, as yet, and no comprehensive framework that can encompass and account for all aspects related to (the annotation of) modality. Indeed, we believe that a comprehensive scheme for the annotation of modality needs to fulfill a set of requirements, and our proposed approach manages to obey them: (i) general flexibility (validity for all approaches); (ii) exhaustiveness (ability to encompass all

[5] Annotation of features following specific guidelines is underway for this part of the pilot study. Preliminary agreement over triggers' features show a Kappa of 0.72 at 10% overlap and 0.83 at 100% overlap, so that it indeed seems wise to allow for more aligned constructions while still preserving reasonable agreement.

Overlap	Prec	Rec	F1	TP/FN/TOT
10%	0.82	1.00	0.90	18/4/22
20%	0.82	1.00	0.90	18/4/22
30%	0.82	1.00	0.90	18/4/22
40%	0.82	1.00	0.90	18/4/22
50%	0.77	0.95	0.85	17/5/22
60%	0.77	0.95	0.85	17/5/22
70%	0.77	0.90	0.83	17/5/22
80%	0.77	0.81	0.79	17/5/22
90%	0.64	0.67	0.65	14/8/22
100%	0.41	0.43	0.42	9/13/22

Table 2: Alignment agreement for *targets* with varying amounts of overlap, with trigger alignment fixed at 10%. Judge a is taken as reference in indicating precision, recall, TPs and FNs.

Overlap	Precision	Recall	F1
10%	0.76	0.62	0.68
20%	0.76	0.62	0.68
30%	0.76	0.62	0.68
40%	0.82	0.68	0.74
50%	0.82	0.68	0.74
60%	0.82	0.68	0.74
70%	0.74	0.68	0.71
80%	0.66	0.68	0.67
90%	0.47	0.47	0.47
100%	0.26	0.25	0.25

Table 3: Alignment agreement for *constructions*, with trigger alignment fixed at 10%, and varying overlap constraints for targets. Judge a is taken as reference in indicating precision and recall.

specific schemes, which have to be interpreted within the larger scheme); (iii) constrained freedom (the scheme has to offer a wide set of possibilities among which only some are realised in a given scheme; the way things are realised is fixed, but the choice of what to realise is free); (iv) a shared abstract syntax for the annotation scheme; (v) a shared semantics for values; (vi) shared practices for the annotation procedure. The rather successful agreement over the identification of constructions – which we have evaluated through a rigorous alignment protocol – shows that in spite of freedom and flexibility, the scheme has a strong potential for implementation. Further evaluation of properties and features in underway, as well as further tests on yet other languages. The annotation tool that we have used is freely available and so are the annotation schemes, with a view to provide as much shared material as possible.

7. Bibliographical References

Luciana Beatriz Ávila, Amália Mendes, and Iris Hendrickx. 2015. Towards a unified approach to modality annotation in portuguese. In Malvina Nissim and Paola Pietrandrea, editors, *Proceedings of the IWCS Workshop on Models for Modality Annotation*, London.

Kathryn Baker, Michael Bloodgood, Mona Diab, Bonnie J. Dorr, Ed Hovy, Lori Levin, Marjorie McShane, Teruko Mitamura, Sergei Nirenburg, Christine Piatko, Owen Rambow, and Gramm Richardson. 2010. SIMT SCALE 2009 - Modality Annotation Guidelines, Technical Report 004. Johns Hopkins University, Baltimore, MD.

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

Tullio De Mauro. 1993. *Lessico di frequenza dell'italiano parlato*. Etas.

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 CoNLL '10: Shared Task*, pages 1–12, Stroudsburg, PA, USA.

Dylan Glynn and Karolina Krawczak. 2014. Operationalisation and robust manual annotation of non-observable usage-features. an exploratory study in english and polish. In *Workshop on Modal Categories at EMEL'14*, Universidad Complutense de Madrid.

Iris Hendrickx, Amália Mendes, and Silvia Mencarelli. 2012. Modality in text: a proposal for corpus annotation. In Nicoletta Calzolari et al., editor, *Proc. of LREC'12*, Istanbul, Turkey. ELRA.

Frederic Landragin, Thierry Poibeau, and Bernard Victorri. 2012. Analec: a new tool for the dynamic annotation of textual data. In *Proc of LREC'12*, pages 357–362.

Ben Medlock and Ted Briscoe. 2007. Weakly supervised learning for hedge classification in scientific literature. In *ACL*, volume 2007, pages 992–999. Citeseer.

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

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

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

Malvina Nissim, Paola Pietrandrea, Andrea Sanso, and Caterina Mauri. 2013. Cross-linguistic annotation of modality: a data-driven hierarchical model. In *Proc. of the 9th Joint ISO - ACL SIGSEM Workshop on Interoperable Semantic Annotation*, pages 7–14, Potsdam, Germany, March. ACL.

Paola Pietrandrea. submitted. Epistemic constructions at work: A corpus-driven study on spoken italian dialogues. *Journal of Pragmatics*. `https://www.researchgate.net/publication/296484341_Epistemic_constructions_at_work_A_corpus_study_on_Italian_spoken_dialogues`

Jan Pomikálek, Pavel Rychlỳ, Adam Kilgarriff, et al. 2009. Scaling to billion-plus word corpora. *Advances in Computational Linguistics*, 41:3–13.

Victoria L Rubin. 2010. Epistemic modality: From uncertainty to certainty in the context of information seeking as interactions with texts. *Information Processing & Management*, 46(5):533–540.

Liliana Mamani Sanchez and Carl Vogel. 2015. A hedging

annotation scheme focused on epistemic phrases for informal language. In Malvina Nissim and Paola Pietrandrea, editors, *Proc. of the IWCS Workshop on Models for Modality Annotation*, London.

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

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.

György Szarvas, Veronika Vincze, Richárd Farkas, György Móra, and Iryna Gurevych. 2012. Cross-genre and cross-domain detection of semantic uncertainty. *Computational Linguistics*, 38(2):335–367.

Paul Thompson, Giulia Venturi, John McNaught, Simonetta Montemagni, and Sophia Ananiadou. 2008. Categorising modality in biomedical texts. In *Proc. of the LREC 2008 Workshop on Building and Evaluating Resources for Biomedical Text Mining*, pages 27–34.

Veronika Vincze, György Szarvas, György Móra, Tomoko Ohta, and Richárd Farkas. 2010. Linguistic scope-based and biological event-based speculation and negation annotations in the Genia Event and BioScope corpora. In Nigel Collier et al., editor, *Proc of the Fourth International Symposium for Semantic Mining in Biomedicine, Cambridge, UK*, volume 714 of *CEUR Workshop Proceedings*.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2-3):165–210.