

# MPQA Gate Annotation – Annotation Details

May 4, 2008

## 1 Annotation Sets and Types

There are three annotation sets that will be in the Annotations Sets frame of GATE:

1. Default annotations
2. MPQA annotations
3. Original markups

The annotations types that you will work with for this task are those under MPQA annotations.

Below, the MPQA annotation types are listed with brief descriptions of their possible features.

### agent

1. *id* - A unique identifier assigned by the annotator to the *first* meaningful and descriptive reference to an agent. This id is case sensitive. When annotating a later mention of the same referent as an agent, there is no need to re-specify the id.
2. *nested-source* - Added to an agent annotation when the agent reference is the source of a private state/speech event. It is a list of agent ids beginning with the writer and ending with the id for the immediate agent being referenced. For instance, w,smith,miller would capture the idea that the writer quotes somebody called Smith who in turn quotes somebody called Miller.
3. *nested-target* - In the present version of the scheme, this field is not used. Targets are specified in attitude annotation frames.
4. *agent-uncertain* - Use when you are uncertain as to whether or not the agent is the correct source of a private state/speech event.

Possible values: somewhat-uncertain, very-uncertain

**NOTE:** Please see [Annotating Agents](#) for more detailed instructions and examples.

## expressive-subjectivity

1. *es-uncertain* - Use when you are uncertain as to whether or not the word or phrase you are annotating is an expressive-subjective element.

Possible values: somewhat-uncertain, very-uncertain

2. *intensity* - The strength of the expressive-subjective element.

Possible values: low, medium, high, extreme

3. *nested-source* - Agent that is the source of the private state indirectly indicated by the expressive-subjective element. It is a list of agent ids beginning with the writer and ending with the id for the immediate agent that is the source.

4. *nested-source-uncertain* - Use when you are uncertain as to whether or not the agent is the correct source for the private state indirectly indicated by the expressive-subjective element.

Possible values: somewhat-uncertain, very-uncertain

5. *polarity* - Attribute for marking the polarity of the expression, in context, according to the nested-source.

Possible values: negative, positive, both, neutral, uncertain-negative, uncertain-positive, uncertain-both, uncertain-neutral

6. *attitude-type\** - This attribute is no longer used now that attitudes are being explicitly annotated. It was used if the expressive-subjective element was expressing a negative or positive attitude, feeling, evaluation, or emotion.

Possible values: negative, positive, other

## direct-subjective

1. *annotation-uncertain* - Use when you are uncertain if, in context, the word or phrase expresses a direct private state/speech event.

Possible values: somewhat-uncertain, very-uncertain

2. *attitude-link* - This contains a list of the ids of all attitudes that are associated with the private state expressed by the direct-subjective

3. *expression-intensity* - Strength of the private state being expressed by the direct-subjective expression. To give you an idea, 'said' is neutral, 'thinks' is low, 'criticized' or 'fears' is medium, and something like 'blasted' in the verbal sense is probably high.

Possible values: neutral, low, medium, high, extreme.

4. *implicit* - Add this feature when you annotate a (zero) span of text for an implicit speech or thought. For example, there may be quoted speech without a "said" where the speaker is implicit from the previous sentence. In this case, make the first quote or word at the beginning a direct-subjective and use this feature.
5. *insubstantial* - Use when the private state/speech event is not significant or not particular, based on the criteria for significant and particular in the annotation instructions. Type in all criteria that it fails to pass: c1 and/or c2 and/or c3.
6. *intensity* - The overall strength of the private state being expressed. Think of this as the union of the intensity of the expressions plus the strength of the private state being expressed by the expressive- subjective elements.

Possible values: neutral, low, medium, high, extreme

7. *nested-source* - Agent that is the source of the private state/speech event. It is a list of agent ids beginning with the writer and ending with the id for the immediate agent being referenced.
8. *polarity* - New attribute for marking the polarity of the expression, in context, according to the nested-source.

Possible values: negative, positive, both, neutral, uncertain-negative, uncertain-positive, uncertain-both, uncertain-neutral

9. *subjective-uncertain*- Use when you are uncertain, in context, whether the word or phrase ought not to be treated as an objective-speech event.
10. *attitude-toward\** - This feature is no longer used. It was used for cases where there was a negative/positive attitude being expressed and the attitude was being directed toward an agent. In that case this feature received the id of the relevant agent.
11. *target-speech-link\**- Deprecated.

## objective-speech-event

1. *annotation-uncertain*- Use if you are unsure that the word or phrase is really used to refer to a speech event.
2. *implicit*- Add this feature when you annotate a (zero) span of text for an implicit speech. For example, there may be quoted speech without a "said" where the speaker is implicit from the previous sentence. In this case, make the first quote or word at the beginning an objective-speech event and use this feature.

3. *insubstantial*- Use when the speech event is not significant or not particular, based on the criteria for significant and particular in the annotation instructions. Type in all criteria that it fails to pass: c1 and/or c2 and/or c3.
4. *nested-source*- Agent that is the source of the speech event. It is a list of agent ids beginning with the writer and ending with the id for the immediate agent being referenced.
5. *objective-uncertain*- Set this feature if you are unsure whether the speaking event might not better be treated as a direct-subjective.
6. *target-speech-link\**- Deprecated.

## target

1. *id*- a unique id for this target. Note that even if this very same target reoccurs as a target elsewhere, give it a unique target id every time.
2. *target-uncertain*- Use if you are unsure that the selected word or phrase really is the target of the attitude to which you related it.

## attitude

1. *attitude-type*- The specific attitude subtype that you recognize. The possibilities consist of the following set: agree-neg, agree-pos, arguing-neg, arguing-pos, intention-neg, intention-pos, other-attitude, sentiment-neg, sentiment-pos, speculation
2. *attitude-uncertain*- Use when you are uncertain about the presence of an attitude, or when you are not sure what the subtype of the attitude is.
3. *contrast*- Set to yes if the attitude conveyed arises as part of a contrast between two situations.
4. *id*- A unique id for this attitude.
5. *inferred*- Set to yes if the attitude you're marking is just an inference. (E.g. it would apply in the case of sentiment-neg towards the target Chavez in the oft-cited example "People are happy that Chavez fell")
6. *intensity*- This feature captures the strength of the attitude expressed.

7. *repetition*- Use if the attitude is conveyed through the use of repetition.
8. *sarcastic*- Use if the attitude that is being conveyed is sarcastic.
9. *target-link*- A list of ids of the target spans that are associated with this attitude.

## inside

In the course of normal annotation, you should almost never need to edit these annotations. When you adjust sentence splits, you might have to merge or extend insides.

If you create new inside labels, make sure that the nested-source is correct. It almost always has to be 'w' for writer.

1. *nested-source*- the agent that is the source of the sentence's content
2. *comment*- Use this feature if, for a given sentence, you want to record a comment about something you did or didn't annotate in the sentence.
3. *error\**- Deprecated.
4. *inside-uncertain\**- Deprecated

## split

Indicates the sentence and paragraph splits.

## 2 Annotating Agents as Sources

In these annotations, there is one role that an agent can fill, namely that of being the source of a private state or speech event.

In an article, an agent may be referenced any number of times, and may be a source for any number of speech events or private states.

Consider the following sentence.

(1)China said on Tuesday (2)a U.S. State Department report that accused (3)Beijing of suppressing religious freedom was full of lies.

In this example, there are two agents that we are interested in: China and a U.S. State Department report. Also, there are two references to the agent China. These are the (1)China and (3)Beijing spans above.

The annotations for the phrases (1) China, (2) a U.S. State Department report, and (3) Beijing are explained below.

**AGENT:** China

The sentence begins, "China said." Here China is the agent (according to the writer) that is the source of the speech event indicated by 'said'. The span, 'China', is annotated as an agent with the following features:

- (a) **id=china**
- (b) nested-source=writer,china

Because this is the first meaningful reference to China, the agent, the 'China' span annotation is assigned an identifier (id=china) that will be used to refer to the agent China in any annotation throughout the document. **You should NOT add the feature, id, to any another other annotation referencing the agent China, anywhere else in the document.**

The nested-source feature of the 'China' span annotation indicates that China is the source of the speech event, 'said'.

**AGENT, target:** a U.S. State Department report

The second agent to annotate is the span for the U.S. report.

Notice that the U.S. report is not only a source for the speech event 'accused', but it is also the target of the negative emotions (attitude) of China. China thinks the report is "full of lies." So the report is both a source and a target. In the agent annotation, capture only the function of the agent as a source by using the nested-source feature. Also, enter the agent id in the nested-source feature of the DSE 'accused'.

Then add a target label on the span 'a U.S. State Department report', give it an id, and link it to the writer's negative sentiment attitude.<sup>1</sup>

When indicating that an agent annotation is a nested-source, we maintain the nesting. The report, according to China, according to the writer, is accusing.

Not that in target annotation frames, there is no need to display the nesting of sources. For instance, the fact that the report is full of lies according to China, according to the writer can be derived by 'going upstream' from the target 'US State Department report' to the negative-sentiment attitude that it links to, and then on to the DSE 'said' that the attitude links

---

<sup>1</sup>Note that this policy is new. Previously, the way to deal with a situation where an entity serves as an agent and as a target of two different private states/attitudes was the following: in the agent annotation frame, the nested-target feature, would have been used to capture the fact that the referent of the agent phrase also serves as a target. Thus, in the older way of doing things, the annotation frame for "a U.S. State Department report" would have looked as follows:

- (a) id=report
- (b) nested-source=writer,China,report
- (c) nested-target=writer,China,report

This is not the current practice! Follow the practice given above in the main text.

to. Inside the annotation frame for ‘said’, we can look at the nested-source feature to determine the nesting.

**target:** Beijing

This is the second reference to China. China(Beijing) is being accused by the U.S. report of suppressing religious freedom, so it is the target of a negative attitude from the US.

Thus, we have to create a target label for Beijing, give it a unique id, and link it to the negative sentiment attitude that belongs to the DSE ‘accused’ and has the U.S. State Department report as its source.<sup>2</sup>

### 3 Rules for Annotating Agent Spans

1. Every unique agent referred to in the text should be assigned ONLY ONE identifier. In other words, out of all agent spans in a document that refer to the U.S. human rights report, only one of them will have the feature, id.
2. Note that this policy is different from that for targets: if the same entity occurs as a target multiple times in a text, it will be assigned a unique id on each occasion.
3. Agent ids are case sensitive! If you give an agent an id=AbCdEf then you must type AbCdEf as the id for that agent every time you reference that agent in a nested-source, nested-target, etc.
4. The id feature should be assigned to the first descriptive reference to the agent. Finding this reference is usually clear-cut but in some cases it’s harder because the information that helps one to identify the agent referent is more distributed. Consider this example:

- (a) So much for President Bush’s effort to repair his legacy on global warming at least when it comes to one German official with a flair for sloganeering.

In a statement released today, Environment Minister Sigmar Gabriel described Mr. Bush’s speech on Wednesday as “disappointing.”

---

<sup>2</sup>The way Beijing is annotated is also a departure from previous practice. In the older way of labeling, the span ‘Beijing’ was annotated as an agent with the following feature:

- (a) nested-target=writer,China,report,China

The last id on the list of id for the nested-target was intended to tell us which agent the span is referring to (China). Further, the nesting indicated by the nested-target was intended to show according to whom China is the target of a negative attitude.

As pointed out above, we no longer use the nested-target feature in agent annotation frames. Use separate target-labels.

In the second sentence, where the DSE “described” occurs, the relevant agent phrase is “Environment Minister Sigmar Gabriel”. The question is whether one should consider the previous reference to “one German official with a flair for sloganeering” as an earlier descriptive reference. Here it seems acceptable to treat only the second mention where the person is identified with their office and name as fully descriptive. The mention in the first sentence would thus not have to be marked as an agent.

5. When annotating a span of text that references an agent, label the entire noun phrase that is part of the reference. Thus, in the previous example, mark “Environment Minister Sigmar Gabriel” rather than only “Sigmar Gabriel”.

## 4 A Strategy for Annotating in GATE

Before you begin annotating, it will make your life easier if you start by checking the boxes for the agent, on, expressive-subjectivity, and split annotations. It will also make your life easier if you sort the annotations by their starting byte, so that you can more easily keep track of your annotations.

The basic recommendation is to proceed sentence by sentence and to perform the steps below for each sentence. Of course, they are meant only as a recommendation and you should do whatever works best for you.

1. First, look at annotations that pertain to the writer of the document as a whole.
  - (a) Find and annotate all expressive-subjectivity for the writer. Edit the annotations, setting the nested-source and intensity features. Set the polarity-type feature if, in context, the expressive-subjective element is expressing a negative or positive attitude, or expresses a combination of both positive and negative attitudes.
  - (b) Use the writer’s expressive-subjectivity annotations from the previous step to help determine if the sentence-level private state annotation for the writer should be a DSE or an OSE. You need to make a change only if the annotation for the writer is DSE since by default an OSE-annotation is provided. If you do change the type from OSE to DSE, also remember to set the other appropriate features such as intensity. Of course, when the DSE is implicit, you do not need to specify a value for the feature expression-intensity.
  - (c) Apply attitude and target labels as per Theresa’s instructions. Make sure that for each attitude you specify at least
    - i. id
    - ii. attitude-type
    - iii. intensity
    - iv. target-link
  - (d) Label the appropriate target span and give it an id. If you are unsure that the span really functions as the target of the attitude to which you are linking the target in question, then set the target-uncertain feature.

- (e) Also make sure that after completing an attitude annotation, you enter its id into the attitude-link field of the relevant DSE annotation frame.
  - (f) If the attitude you marked was an inferred one, don't forget to set the inferred feature to "yes".
2. Turn to the more deeply embedded nested-sources in the sentence. These will be typically mentioned overtly but might be implicit.
- (a) Identify in the sentence all other direct mentions of private state and speech events that meet the criteria for annotation. That is, find OSEs and DSEs.
  - (b) For every private state/speech event that you identified in the previous step, annotate
    - i. the span of text that evokes the private state/speech event
    - ii. the span of text that refers to the agent that is the source
    - iii. any spans of text that are expressive-subjectivity attributed to the source of the private state/speech event
  - (c) Edit the agent annotations that you just added, providing an id if needed, and setting the nested-source feature. Make sure that if the agent appears as a source for the first time, you also find the first descriptive reference to that agent in the text and mark it, giving it its initial id.
  - (d) Edit the expressive-subjectivity annotations that you just added, setting the nested-source and intensity features. Set the polarity feature if in context, the expressive- subjective element is expressing a negative or positive attitude.
  - (e) If the OSE or DSE is implicit, set the implicit feature to "true". A common situation in which an embedded source and their private state expression go unexpressed is when a sentence continues a quote as in the second sentence of this example:
    - (1) "That is a pessimistic assessment, but it may be realistic," he wrote in an email. (2) "Look, for example, at the E.U. where, ... total E.U. emissions are now, once again, inching back up."
  - (f) Specify the nested-source for OSEs and DSEs.
  - (g) If you are annotating a DSE, specify the expression-intensity and intensity features. (Omit expression-intensity if the DSE is implicit.)
  - (h) If the source of a DSE is expressing any kind of attitude, you need to start adding attitude labels and, where appropriate, target labels, and link them appropriately.
  - (i) If an attitude you marked was inferred, remember to set the inferred feature to "yes".
3. Things to keep in mind
- (a) Often, when you have an ESE marked on a sentence you will also have an attitude.

- (b) We can mark attitudes that are inferred—think of the classical case of “People are happy that Chavez fell”.
- (c) A single private state expression may have multiple attitude annotations associated with it.
- (d) If you mark an attitude as inferred, then you should also have another attitude present that is not inferred.
- (e) Many annotation frames allow you to mark uncertainty. If you really are uncertain, use the appropriate fields.

## 5 Adding Comments To Insides

Although we are not working with the insides of the private states and and speech events at this time, the ‘inside’ annotations for the level of the writer, which span sentences, were included during document preparation. These ‘inside’ annotations have a comment feature.

Use this feature if, for a given sentence, you want to record a comment about something you did or didn’t annotate in the sentence. You may edit the ‘inside’ annotation for the writer for that sentence and add the comment feature. Type in your comment as the value for the feature.

## 6 Handling Bad Sentence Splits

Note that the below differs from previous policy. Also note that the discussion below is most relevant to the internal concerns of the Pitt annotation group.

Before you annotate a new document, check out the sentence splits. Note that there are two kinds of splits, the default GATE\_Splits that come from the text processing platform, and the MPQA splits. The preprocessing done in GATE sets the MPQA splits to be the same as the GATE Splits.

Splits could be bad in one of two ways: their extent is too small or large, or they are in the wrong place, where wrong place typically means that they need to be deleted, for instance because a split got introduced because of an abbreviation ending in a period.

When you modify splits, change both the GATE and the MPQA splits. (Since I am not sure at this point whether one or the other type of split is crucial to automatic systems let’s adjust both.) Also let’s adjust the associated GATE\_Sentence labels and the MPQA\_inside labels.

For instance, if you had a split after “Mr” in :

Mr. Bean ... You’ll just have to love him!

you would want to remove it (both the GATE Split and the MPQA split) and then you need to merge the two MPQA insides that cover “Mr” and “... You’ll just have to love him!”. Likewise, you need to merge the two GATE\_Sentences over the same spans.

## 7 Checking Your Annotations

When you have finished annotating a document, you need to check your annotations. I suggest that you do this after taking a break, possibly until the next day.

Double check that:

1. No ids are missing in the agent, dse, ose, and expressive-subjectivity annotations.
2. Ids for a given agent match wherever that agent is referred to in annotation features. (Check for typos.)
3. Make sure that polarity and intensity are specified where needed.
4. You didn't miss any annotations for private state/speech events or expressive subjective elements.

A good way to check your annotations is to (a) sort the list of annotations in the Annotations frame by starting byte, (b) select the first annotation, (c) step down through the annotation list using the arrow key. As you step through the annotations in this fashion, the non-zero-span annotations will flash when they are selected. Check one annotation at a time.

Alternatively, try running the MPQA Annotation Checker. The instructions are at: <http://www.cs.pitt.edu/mpqa/opinion-annotations/gate-instructions/checkerinstr.html>.

Last updated May 4, 2008