Meta-Knowledge Annotation of Bio-Events

Annotation Guidelines

Paul Thompson, Raheel Nawaz, John McNaught and Sophia Ananiadou

School of Computer Science, University of Manchester,
UK
{paul.thompson, john.mcnaught, sophia.ananiadou}@manchester.ac.uk
nawazr@cs.man.ac.uk

Contents

1 Introduction and Background

If a user wishes to search for relevant information located within biomedical documents, the usual method is to enter keywords into a search engine. However, such searches normally return a large number of documents, many of which are likely to be irrelevant.

Assume that the user wishes to find instances of positive regulations involving the protein *narL gene product.* He may enter the search terms "*narL gene product"* and *activate,* since instances of positive regulations are often described using the verb *activate.* Although his goal is to find documents where these search terms are related to each other in a specific way, the problem is that normal search engines do not take account of relationships between search terms, and may even return documents where the 2 search terms are each located in a separate sentence.

Text mining systems help to cut down on the amount of time that users have to spend sifting through irrelevant documents. This is facilitated by providing the user with the means to formulate more structured queries, which ensure that only those documents containing the required type of knowledge are returned by the search. Using a text mining system, the user can specify that he wishes to find all instances of positive regulations, where *the narL gene product* is the instigator of the regulation. It is not necessary to worry about exactly how the regulation is expressed in the text, e.g., which verb is used.

Although text mining systems providing functionality such as the above have already been developed, what they often lack is a means to distinguish between definite facts and other types of interpretations. For example, a text mining system may retrieve the following fact in response to the query above:

(S1) The narL gene product activates the nitrate reductase operon

Sentence (S1) can fairly certainly be interpreted as describing a definite fact. However, compare this to sentence (S2):

(S2) Our results suggest that the narL gene product activates the nitrate reductase operon

In (S2), the first part of the sentence projects a rather different interpretation to the information described by the verb *activates*, i.e., it is a somewhat tentative interpretation/analysis of results, which should certainly not be interpreted as a definite fact.

The ability to distinguish between different interpretations of information can be important, e.g., a biologist may want to search a collection of documents to isolate descriptions of *new knowledge* (e.g., experimental observations and confident analyses of results) from other types of knowledge (e.g., descriptions of well-established knowledge, hypotheses, etc.). This could be useful, for example, in maintaining an up-to-date database of biological interactions. If the isolation of new

knowledge from other types of knowledge can be carried out automatically, this can potentially save the user a large amount of time.

In order to produce systems that can distinguish different interpretations of information, we need to undertake a task called *annotation.* This involves reading texts and identifying and marking (annotating) the different ways in which information relating to the interpretation of knowledge (which we term *meta-knowledge*) can be expressed in texts. The text mining system can then learn to generalize from the annotated examples (using a computer algorithm), in order to be able to assign interpretation information to previously unseen examples. This annotation process is the subject of this document.

1.1 Background to the Task –Searching for Relevant Information

Complex, structured queries such as those introduced above must be matched against structured representations of the biological knowledge that occurs in documents. Text mining systems need to be able to analyse texts in order to locate this biological knowledge and produce structured representations from the unstructured text. These structured representations of knowledge are called *events.* A number of collections of documents (called corpora) contain event annotations. These have been produced by domain experts, in order to allow text mining systems to learn how to recognise relevant events within texts. The meta-knowledge annotation introduced above will be carried out for individual events within these event-annotated corpora. This will provide the necessary information to train systems which not only recognise events, but can also determine automatically how those events should be interpreted.

In this section, we firstly look more closely at why events and event-based searching are needed, by examining the more usual keyword searches, and highlighting their pitfalls. We then move on to look at an example of an event, and how searching using events can be more powerful and can retrieve more focussed results than are possible using keyword searches

1.1.1 Keyword Searching and its Problems

It is often necessary for biologists to search the literature for relevant information. For example, a particular user may be interested in discovering the types of things that are positively regulated by a particular protein, e.g. *the narL gene product.* A sentence such as (S1) would provide the type of information that is sought:

(S1) The narL gene product activates the nitrate reductase operon

In other words, one type of sentence that would help the user to locate the information they require would be one in which *The narL gene product* is the grammatical subject of a verb which describes a positive regulation (such as *activate*). In such a sentence, the grammatical object of the verb (i.e., *the nitrate reductase operon* in the above example) will provide the information that is sought.

As mentioned above, using a search engine such as *Google* or *PubMed* would involve entering keywords and phrases such as *"narL gene product"* and *"activate"*. Although a search carried using these terms is highly likely to retrieve relevant

documents, it is just as likely to retrieve a large number of documents that are not relevant.

Keyword searches such as the above can be problematic for a number of reasons, and can retrieve many irrelevant documents as well as relevant ones. For example:

- Searching for *The narL gene product* and *activate* as separate search terms does not guarantee that they will be grammatically related to each other in the text in the way specified above. The search terms may not even occur within the same sentence.
- Searching using a single quoted search term, e.g., "*The narL gene product activates"*, to ensure that the verb occurs next to the protein in the text, is also not sufficient. The set of documents returned by such a query is likely to be smaller and more relevant than if using separate search terms. However, many relevant documents could also be missed, due to the large number of potential variations in the way that the positive regulation can be expressed in text. Some similar phrasings of the sentence (1) would include "*The narL gene product* ***is known to*** *activate the nitrate reductase operon.", "The narL gene product* ***rapidly*** *activates the nitrate reductase operon", "The nitrate reductase operon* ***is activated by*** *the narL gene product".*
- Positive regulation events may be described by a number of different verbs and nouns other than activate e.g. *increase, affect, effect*

In short, retrieving all relevant documents using simple keyword searches can be rather time consuming, and will often require a number of separate searches to be carried out, and much sifting of the documents returned in order to distinguish those documents that are relevant to the query.

1.1.2 Events and Event-Based Searching

Text mining technology can help greatly in searching for information, both to giving extra power to the searching mechanism, thus reducing the number of separate searches that have to be carried out, as well as increasing the relevance of the results that are returned by the search.

Unlike traditional search engines, text mining systems do not simply view documents as sequences of words, but rather they try to *structure* this information automatically, and try to find relationships between words and phrases within sentences. These structures are called *events* and the automatic process is called *event extraction.*

A possible structured representation of the event described in sentence (S1) would be the following:

EVENT_TYPE: *Positive_Regulation*
EVENT_TRIGGER: *activates*
CAUSE: *The narL gene product* (PROTEIN)
THEME: *the nitrate reductase operon* (OPERON)

The main features of this representation are as follows:

- EVENT_TRIGGER – a word or phrase around which the event is "organized" in the text. This is often a verb (in this case *activates*) or nominalized verb (a noun with a verb-like meaning, such as *transcription* or *activation)*
- EVENT_TYPE - The event is assigned a type from a fixed set of possible values that characterise different types of events in biomedical texts. The event type abstracts away from the actual verb used to describe the event in the text.
- Event participants – Each event has one or more participants. These are generally entities (e.g. genes, proteins, organisms, etc.) that play a part in description of the event. Each participant is separately identified and assigned the following information:
    - Semantic role – a label that characterizes the contribution of the participant towards the description of the event. The labels used are rather general, as they are intended to be applicable to all events in biomedical texts. The following roles are used in the description above.
        - CAUSE – participant responsible for the event occurring
        - THEME – participant affected by or during the event
    - Named Entity (NE) type – a label that characterizes the type of biological entity that the event participant represents (e.g. PROTEIN). Again, these types are chosen from a fixed set of values.

The automatic extraction of such events from texts allows searches to be carried out on these structures themselves, rather than using keyword searches on the unstructured text. The event structure abstracts from the exact wording in the text, meaning that searches over events can specify the following:

- Event types (e.g. *Negative_regulation, Binding*) instead of precise verbs or nominalised verbs used to describe the event
- Restrictions on the event participants in terms of:
    - Semantic roles specified by the event (e.g., CAUSE, THEME)
    - Values of particular roles, which could be specified as either:
        - Keywords when searching for specific values (e.g., *narL gene product*)
        - NE types for a more general search (e.g. events where the CAUSE is any entity of type PROTEIN)

Thus, the user has a choice about how general or specific to make their query. NE and event types are often arranged into a hierarchy, giving the use even more control over how general or specific their search will be.

As event-based searching allows users to be more precise about the type of information they are looking for, the set of results is better aligned with the users requirements, i.e., the results are more focussed, and contain fewer irrelevant documents than simple keyword searches. The results are also more concise than those returned by a traditional search engine, showing only the relevant events, or the sentences from the documents in which the relevant events are contained, rather than complete documents.

In more complex sentences, it is possible for multiple events to be present, and it is also possible for the participant of a particular event to be another event. Consider example (S3).

(S3) We found that Y activates the expression of X

Here, the "main" event in the sentence, i.e., the one which is triggered by the verb *activates*, has a similar structure to the event in sentence (S1), except that the *THEME* of the event (i.e. *the expression of X*) is not a simple entity, so how do we deal with it?

EVENT_TYPE: *Positive_Regulation*
EVENT_TRIGGER: *activates*
CAUSE: *Y*
THEME: *?*

We actually treat this THEME as being a separate event, as it can be seen as having its own structure, with the type *GENE_EXPRESSION* and the *THEME* of X. Note that is not necessary for both CAUSE and THEME to be specified for all events. To deal with the fact that this second event is a participant of the first, we assign the unique identifiers *E1* and *E2* to the events. Figure 1 shows the full representation of these 2 events.

Using this notation, the biological knowledge contained in a document can be represented a set of events, some of which will be "nested" within each other.

We refer to E2 as a primary event, and E1 as a secondary event. E2 conveys the main information, whilst E1 can be seen as providing supporting information – it is not a complete or "interesting" piece of knowledge in itself. It is often (but not exclusively) the case that primary events have event triggers that are verbs, whilst secondary events have triggers that are a special type of noun with a verb-like meaning called *nominalised verbs.* The noun *expression* is an example of one of these, with a meaning similar to the verb *express.* Other examples would include *transcription* (from the verb *transcribe*) and *regulation* (from the verb *regulate*).

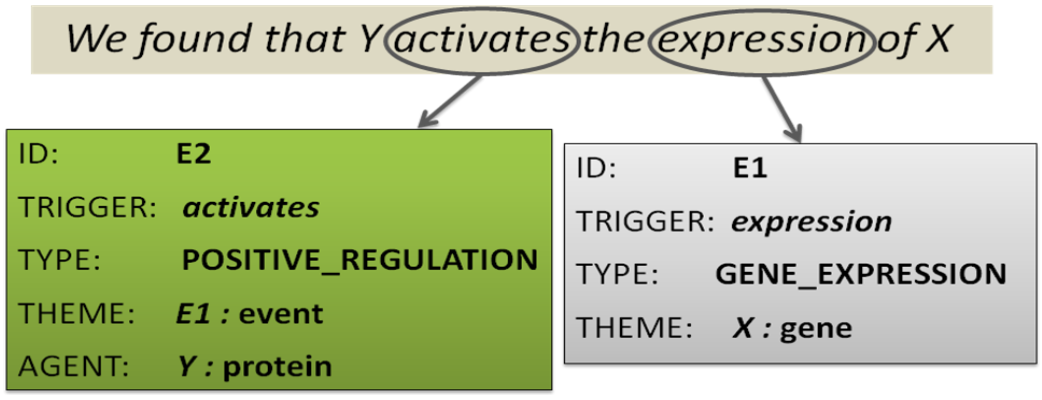

Figure 1 – Event Representation Example

1.2 Need for Meta-Knowledge Annotation

Text mining systems are normally trained to recognise events by learning from annotated examples. That is to say, a corpus of document (called a *corpus*, plural *corpora*) are annotated with events by human domain experts. The event annotation process often involves:

- Locating the event trigger
- Assigning a type to the event
- Identifying the participants of the event
- Assigning roles and NE types to these participants

In the biomedical field, a number of such annotated corpora already exist, making it possible to train systems to recognize events and their participants. However, information about the *interpretation* of the events (i.e., meta-knowledge) is often missing from the annotation, or it is not dealt with in a satisfactory way.

Some examples of meta-knowledge that we consider to be important include the following:

- Is the event negated?
- Is the event stated with complete certainty, or is there some degree of uncertainty conveyed?
- Does the event describe well-established knowledge or new knowledge? New knowledge may correspond to direct observations, or an analyses made by the author based on experimental results
- What is the intensity of the event? (e.g. *strong* or *rapid* vs. *weak* or *slow*)

A text mining system that can distinguish between these different types of interpretations can clearly be useful to users. For example, positive and negative events have completely different interpretations. Likewise, it would be useful to present to the user some indication of the reliability of the event, e.g. events explicitly marked as possibly true need to be distinguished from those events which are known to be definite. In a similar way, analyses based on results are less reliable than direct observations. The ability to distinguish between new and well-established knowledge may be useful in applications, such as curating a database of known protein interactions.

In order to allow precise meta-knowledge to be recognized at the level of events, the annotation task described in this document will identify and assign different types of meta-knowledge to each individual event in a document.

1.2.1 Meta-Knowledge Examples

To make the ideas of meta-knowledge introduced above more concrete, let us consider 8 sample sentences, the majority of which contain 2 basic events:

1) A *positive regulation* event where *Y* is the AGENT, and the *expression* event described in 2) is the THEME
2) An event describing a *gene expression*, where *X* is the THEME

Note that, in most cases 1) is the primary event in the sentence, whilst 2) is the secondary event. It is normally the case that most meta-knowledge information expressed in the sentence will apply to the primary event. Often there is no information that allows a specific interpretation to be applied to a secondary event. This is not exclusively the case, although here we concentrate mainly on the interpretations of the primary events in the sentences.

The sample sentences are as follows:

(S3) ***We found*** *that Y activates the expression of X*

(S4) ***We examined*** *the effect of Y on expression of X*

(S5) *These results* ***suggest*** *that Y has* ***no*** *effect on expression of X*

(S6) *Y is* ***known*** *to increase expression of X*

(S7) *Addition of Y* ***slightly*** *increased the expression of X*

(S8) *These results* ***suggest*** *that Y* ***might*** *affect the expression of X*

(S9) ***Significant*** *expression of X was* ***observed***

(S10) ***Previous studies*** *have* ***shown*** *that Y activates the expression of X*

The trigger words for the events are underlined in each of the examples. The *expression* event, which occurs in all sentences, is always indicated by the nominalised verb *expression.* However, the positive regulation event is expressed in a number of different ways, namely using the verbs *activate*, *increase* and *affect*, or the nominalised verb *effect.* The positive regulation event occurs in all sentences, with the exception of (S9).

The emboldened words and phrases in the examples below help to show that the way in which the events should be interpreted can vary considerably. However, current text mining systems will normally treat the events extracted from all the above sentences in an identical way, thus missing important or even vital details about the event. Most of the emboldened words affect the interpretation of the positive regulation event, which is the main event in the sentence. However, in (S9) the interpretation of the expression event is altered.

In sentence (S3) above, the presence of the word *found* shows explicitly that the positive regulation event is backed by evidence, i.e. it is an experimental observation. The word *we* shows that is very likely that event was observed by the authors of the paper as part of the study being described, which would mean that it could be considered as "new" knowledge. No explicit information is specified for the secondary expression event, although we also consider this to be an observation.

The interpretation of the positive regulation event in (S10) is very similar to (S3). The presence of the word *shown* is again an explicit indication that the positive regulation event is an experimental outcome. However, the use of *Previous studies* at the start of

the sentence indicates that these results were originally reported outside of the current paper, and hence the event should not be considered as not "new" knowledge. Once again, there is no explicit information regarding the secondary expression event, but again we would treat this as an observation

Sentence (S6) also contains events with similar interpretations to those in (S3) and (S10). However, the word *known* serves to indicate that the positive regulation event is a well established fact within the field. Whilst (S3) and (S6) can be seen as representing the same type knowledge at some level, in that they both report the event is a definite fact which is backed by evidence, the degree of the "reliability" of the events is subtly different, in that (S3) reports a new experimental outcome rather than well-established knowledge.

Whilst there are subtle differences in the interpretation of the positive regulation events in (S3), (S6) and (S10), they all have in common that the event is presented as without any expression of uncertainty. In this respect, the positive regulation event in (S4) is quite different. Here, the presence of the word "examined" serves to indicate that the positive regulation event is under examination, and so, at least at that point in the text, it is not possible to determine whether or not the event is true. Thus, it would be incorrect for a text mining system to present the positive regulation event in this context as a definite fact or an observation.

In (S8), there is yet a different interpretation of the positive regulation event. In using the word *might*, the author is indicating some amount of speculation towards the truth of the event. Furthermore, the use of the verb *suggests* denotes that the evidence for the author's tentative statement is based on some kind of analysis or inference drawn from results. Such evidence is, by its nature, less reliable than the direct evidence than was stated to be behind the positive regulation events in (S3), (S6) and (S10).

Sentence (S5) is similar to (S8), in that it also uses *suggests* to indicate that the positive regulation event is based on the results of an analysis. However, the conclusion is different: the author concludes is that the positive regulation event *does not* happen, indicated by the use of the word "no". Hence, this is a negative event.

In sentence (S7), the word *slightly* provides explicit information about intensity of the positive regulation. In (S9), there is only one event, i.e. the expression event. Here, this event becomes the primary event in the sentence, even though its trigger in the nominalised verb *expression.* The intensity of the expression event is indicated, i.e., *significant*. The use of the word *observed* in this sentence shows that this expression event corresponds to an experimental observation.

From the above sentences, we can identify at least five important pieces of interpretative information which can be regularly deduced about events, according to the context in which they appear. These types of information modify the default interpretation (i.e. as positive, definite facts) of the events:

1) What kind of evidence is there for the event, e.g. has it been experimentally observed, inferred from experimental results, is a well established fact, or is it a hypothesis whose truth has yet to be determined?
2) How certain is the author about whether the event is true?

3) Is the event positive, or is it negated (through the use of *no*, *not* etc.)

4) What is the intensity or magnitude of the event?

5) What is the source of the information contained within the event? Is it reported in the current paper or another paper?

The level of impact of each piece of contextual information varies from fairly subtle to fairly significant. However, even subtle information can be important, depending on the task being undertaken or the goals of the user. Therefore, we wish to perform annotation which will capture evidence in the text for all of the above types of information The next section provides more details about the annotation scheme we have designed to allow the above types of information to be made explicit.

2 The Annotation Scheme

Based on the types of meta-knowledge highlighted in the previous section, which appear to be most pertinent to the interpretation of bio-events, we have defined a scheme to annotate these within biomedical texts.

At the heart of scheme are 5 meta-knowledge *dimensions*, which are called *Knowledge Type, Certainty Level, Manner, Polarity and Source* (Figure 2). The other boxes in figure 2 show the types of information that have typically previously been annotated for events in biomedical texts. Each of the meta-knowledge dimensions, which are described in detail in the following subsections, corresponds to a particular type of meta-knowledge. The annotation task consists of two main steps, which are further clarified in the subsections below describing the individual dimensions

1) For each event, determining an appropriate value (from a fixed set) for each dimension, based on evidence from the context in which the event occurs (e.g., the sentence in which the event is described, or previous sentences). The type of evidence that is present can vary. Most often, the presence of particular word or phrase in the same sentence is used as the evidence. In other cases, the evidence constitutes another feature of the sentence, or even the position of the sentence within the abstract.
2) If the evidence for the assignment of a value is a particular word or phrase in the same sentence as the event, then this word or phrase is explicitly marked as a "clue", as part of the annotation task.

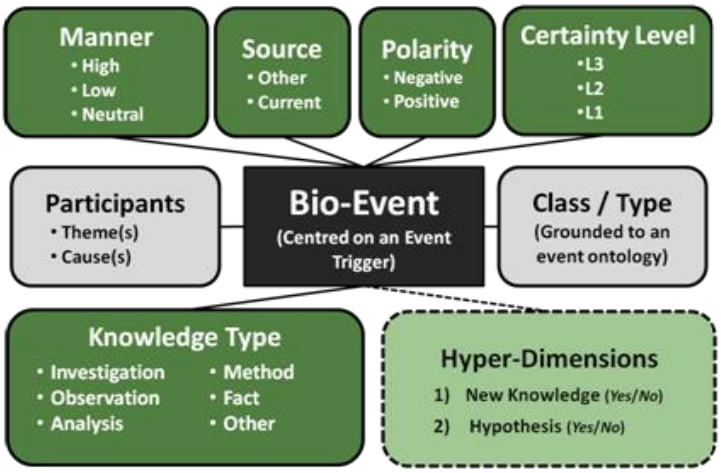

Figure 2. Meta-knowledge annotation scheme

The purpose of the annotation, then, is to discover the different ways in which each value of each dimension can manifest itself as evidence in the text. When we have annotated a large enough set of documents, we can train a system to learn patterns based on these annotations. The trained system will then be able to predict the values of the annotation dimensions for previously unseen events.

In the following sections, we provide detailed information regarding the 5 individual meta-knowledge dimensions. A brief description of each dimension is followed by an enumeration of its possible values, together with some examples. In all of the

examples, the word(s) on which the event is centered (i.e. the trigger word/phrase) are shown using *underlined italics*, whilst the explicit "clue" words which provide evidence for the assignment of a particular value to a dimension are shown using **bold face.**

2.1 Knowledge Type

This dimension corresponds to the general information content of the event. There are six possible values, namely *Investigation, Observation, Analysis, Fact, Method* and *Other*. Most examples given concern primary events. Under normal circumstances, the Knowledge Type of the secondary event is determined on the basis of the Knowledge Type assigned to the primary event, unless there is clear evidence that the secondary event belongs to one of the other Knowledge Types. Further details are given below.

2.1.1 Investigation

Assigned to events that correspond to enquiries or investigations, which have either already been conducted or are planned for the future.

- Evidence – Always indicated through an explicit word or phrase in same sentence as event ***except in titles***. Typical types of evidence include:
  - Verbs in finite form (i.e., showing tense), e.g., *examine, investigate, analyze / analyse, evaluate, study, test, compare, focus* and *explore* etc. Examples (S11-S14) below correspond to such cases.
    - The *Investigation* clue word normally comes before the event trigger, as in (S11 - S13).
    - In the case of passive sentences (e.g. (S14)), the clue word will come after the event trigger
  - Nominalisations of the above verbs (e.g. *investigation, examination, analysis,* etc.) can also indicate investigations (S15)
  - Verbs in infinitive form (i.e., preceded by *to)*. These will normally precede the event-trigger. The verbs that may be used include all of the above, along with some others like *define, ascertain, identify* and *elucidate* etc. An example is shown in (S16).
  - Events in titles can also describe investigations *without the presence of an explicit clue word*. However, this is normally ONLY the case when the title DOES NOT contain verbs, as such titles generally describe topics of investigation rather than definite results (S17 – S18)

    NOTE: Events in titles that DO contain verbs should be treated like other sentences, i.e. an event would only be annotated with the Knowledge Type of *Investigation* if an explicit clue word was present.
- Typical position in text - Towards the beginning of texts, in order to describe the investigation that is going to be carried out.
- Secondary events – If the primary event has the Knowledge Type of *Investigation*, secondary events will normally have the Knowledge Type *Other*. It is possible that the secondary event may be assigned *Analysis,* if it is clearly stated based on an analysis.

- Example sentences:

  (S11) We have **examined** the *effect* of leukotriene B4 (LTB4) on the expression of the proto-oncogenes c-jun and c-fos.

  (S12) We **looked at** the *modulation* of nuclear factors binding specifically to the AP-1 element after LTB4 stimulation.

  (S13) To dissect the molecular basis for the unusual persistent expression of the IL-2 and IL-2-R alpha genes in these IARC 301 T cells, we have **analyzed** the *interactions* of constitutively expressed nuclear proteins with the 5' flanking regions of the IL-2 and IL-2-R alpha genes using both DNase I footprinting and gel retardation techniques.

  (S14) *Activation* of expression of genes encoding transcription factors: c-fos and c-jun was **investigated**.

  (S15) **Analysis** of the *expression* of human I kappa B alpha protein in stable transfectants of mouse 70Z/3 cells shows that ….

  (S16) In order to **define** the *roles* of these two factors, which bind to the same kappa B enhancers, in transcription activation we have prepared somatic cell hybrids between IARC 301.5 and a murine myeloma.

  (S17) Constitutive *activation* of NF-kB in human thymocytes (title)

  (S18) *Processing* of the precursor of NF-kappa B by the HIV-1 protease during acute infection (title)

2.1.2 Analysis

Assigned to events for which the truth value is based on inferences, interpretations, speculations or other types of cognitive analysis. This is in contrast to events in the *Observation* category (see 2.1.3), which correspond to directly observable evidence.

- Evidence – **Always** indicated through an explicit word or phrase. Typical types of evidence include:
  - Verbs (finite forms) or their nominalizations preceding the event-trigger, for example, *show, demonstrate, believe, hypothesize, suggest, indicate, appear, seem, conclude, evidence, assume, presume, identify, define, establish, report, reveal, confirm, verify, identify* (S19 – S21)

    NOTE: These verbs denote differing levels of confidence. For example, while *demonstrate* indicates a confident analysis, *suggest* denotes a more speculative conclusion. Therefore, *suggest* also acts as a marker of the

Certainty Level dimension, and should be annotated as such. See section 2.2 for further details.

- Conjunctions such as *therefore* and *thus* etc. These words provide a link to the previous sentence, and implies that some kind of analysis of the results stated in the previous sentence has occurred in order arrive at the stated event. (S22 – S23)

  NOTE: Conjunctions such as *however* and *whereas* act as markers of contrast, and do not have the same kind of meaning as *therefore* and *thus*. So, they should **not** normally be annotated as Analysis markers

- Certain verbs or nominalizations serving as event-triggers and denoting some form of analysis, for example, *correlate, associate, relate, due to, implicate, attribute,* etc. (S24-S25)

- Modal auxiliaries like *may, might* and *could*, as well as adverbs/adjectives like *probably/probable*, *likely* and *perhaps.* These indicate an uncertainty on the part of the author. As such, they also act as markers of the *Certainty Level* dimension (see section 2.2). As this uncertainty must have been reached through some kind of cognitive analysis, they can be considered as *Analysis* markers, but ONLY if no other *Analysis* words are present in the sentence, e.g., (S26-S27). If a finite form of one of the verbs above is also present (e.g., (S28), where *suggest* is present), then it is this finite verb form that should be annotated as the *Analysis* marker

- Frequency indicators such as *often, frequently, normally* and *occasionally* (if no other *Analysis* words are present in the sentence). These denote an analysis on the part of the author as to the perceived frequency of occurrence of the specified event. (S29-S30)

- Adjectives and adverbs (mostly non-finite verb forms) like *is able to, is capable of, suggestive of, consistent with, judged by* and *potential* etc. These again denote analyses on the part of the author. (S31-S32)

  NOTE: The *Analysis* category should NOT be applied to events where the analysis relates only to relative importance of the Agent of the event, rather than to the truth value of the event. An example would be the following:

  Monocytes and macrophages are important *mediators* of Th1-type responses

  In the above example, there is a positive regulation event with the trigger *mediators.* The word *important* denotes that some analysis has taken place, but this analysis regards the relative importance of the mediators, rather than analysis about *whether* the positive regulation event took place. Other similar words include *crucial, central* etc.

- Typical Position in the Text – Towards the end of the text, constituting analyses/interpretations of observations and results described previously

- Secondary events – Normally *Other*, unless the Cause of the event is clearly *Fact* or *Observation*

- Examples Sentences:

  (S19) These results **indicate** that LTB4 may *regulate* the production of different cytokines by modulating the yield and/or the function of transcription factors such as AP-1-binding proto-oncogene products.

  (S19b) The data **suggest** that differences in functional responses elicited in monocytes by all three factors may be *dependent* on different routes on nuclear signaling employed by the factors.

  (S20) Unexpectedly, our in vivo studies also **demonstrate** that I kappa B/MAD-3 *binds* directly to NF-kappa B p50.

  (S21) We also present **evidence** that IL-6 kappa B binding factor II *functions as a repressor* specific for IL-6 kappa B-related kappa B motifs in lymphoid cells.

  (S22) **Therefore**, an indirect *interaction* occurs between these two sites

  (S23) **Thus**, both NF-kappa B-binding complexes are *needed* for optimal viral transcription.

  (S24) Together, this evidence strongly **implicates** BSAP in the *regulation* of the CD19 gene.

  (S25) Moreover, in human T helper (Th) clones functionally characterized as being of the type 0, type 1 and type 2 (28%, < 1% und 93% CD30+, respectively), the extent of CD30-mediated NF-kappa B activation ***correlated*** with the proportion of CD30+ cells.

  (S26) They bind to the kappa B motifs with different relative affinities that **may** *reflect* their different contribution in the expression of various promoters.

  (S27) The MAD-3 cDNA encodes an I kappa B-like protein that is **likely** to be *involved* in regulation of transcriptional responses to NF-kappa B, including adhesion-dependent pathways of monocyte activation.

  (S28) Taken together, these observations **suggest** that HIV gene expression **may** be *activated* in infected monocytes through interaction of the cells with complement-opsonized particles.

  (S29) Our studies now demonstrate that HTLV-1 Tax activates the recently identified cellular kinases IkappaB kinase alpha (IKKalpha) and IKKbeta, which **normally** *phosphorylate* IkappaB alpha on both of its

N-terminal regulatory serines in response to tumor necrosis factor alpha (TNF-alpha) and interleukin-1 (IL-1) stimulation.

(S30) The activation of transcriptional factor c-Fos/c-Jun AP-1 is essential for normal T cell responsiveness and is **often** *impaired* in T cells during aging.

(S31) In addition, IL-2 **is capable of** *increasing* transcript levels of the p50 gene coding for the p50 subunit of the NF-kappa B transcription factor, whereas mRNA levels of the p65 NF-kappa B gene remained unchanged.

(S32) This increase in p50 homodimers coincides with an increase in p105 mRNA, **suggestive of** a *transcriptional up-regulation* of p50.

2.1.3 Observation

Assigned to events corresponding to direct, observable evidence or findings from experiments.

NOTE: A primary event that is the negation of an observation should still be annotated as *Observation*, as this can still be considered as a finding.

- Evidence
  - Explicit word in the same sentence. Typical clue words are *find, detect* and *observe* etc. (S33-S35)
  - If explicit words are not present, the event trigger verb may provide evidence for the assignment of the *Observation* category, if it is either:
    - in the past tense (S36-S37)
    - in the present tense, and in an appropriate context (see below) (S38)
    - A secondary event that is a participant of a primary event assigned the Knowledge Type of *Observation* (S36)
    - Events in document titles (S40)
- Typical position in text
  - Towards the middle of the text, following descriptions of background facts and knowledge, and descriptions of investigations to be carried out, and before analyses of results.
  - Events in paper titles. Titles tend to describes definite experimental outcomes and results, unless there is any suggestion to the contrary. Therefore, most events in titles that are unmarked by clue words and phrases should be annotated with the *Observation* Knowledge Type.
- Secondary events – Typically, if the primary event is an *Observation*, the secondary event is *Observation.* Exceptions include the following (further details under *Other* in section 2.1.4):

a. When the primary event has been negated, and the semantics of this negated event mean that the secondary event did not happen. In this case, the secondary event should be assigned *Other*

b. When the semantics of the primary event mean that the secondary event did not happen. Examples of such primary event triggers include *inhibit, prevent* and *block.*

- Example sentences:
  - Sentences with explicit clue words

    (S33) It was **found** that lipopolysaccharide *induced* strongly both c-fos and c-jun expression as well as AP1 formation.

    (S34) However, no *loss* of DNA binding activity is **observed**, presumably reflecting the unique C-terminal domain that is distinct from that present in NF-kappa B p65.

    (S35) Constitutive DNA *binding* activity consisting of p50 homodimers was **detected** in nuclear extracts from both cell types.

  - Sentences without Explicit Clues (based on Trigger Verbs):
    - Event trigger verb in past tense – this provides fairly reliable evidence for the assignment of the *Observation* category

      (S36) LTB4 ***increased*** the expression of the c-fos gene in a time- and concentration-dependent manner.

      (S37) Both messages rapidly ***declined*** thereafter

      NOTE: In example sentence (S36), there is a secondary event, whose trigger is *expression.* As the primary event is an observation, we also annotate the secondary event with the Knowledge Type of *Observation*, as we assume that this has also been observed.

    - Event trigger verb in present tense – if an explicit *Observation* clue word or phrase is not present in the sentence, the present tense can be ambiguous between describing an observation or a general scientific fact (see the *Fact* category below). Consider sentence (S38):

      (S38) U937 cells ***express*** both type I and type II IFN receptors

      Taken in isolation, the *express* event in (S38) looks most like a general scientific fact. However, by considering the context of the sentence, it may actually be an observation. Taking account of the position of the sentence within the text is often key to determining the correct category. The following two points indicate general

patterns. However, it is important to note that these are only indicative, and do not always occur.

1) Events occurring in the present tense towards the beginning of a text are most likely to correspond to *Fact*, unless the context changes this interpretation.

2) In abstracts that are written completely in the present tense, there is normally an explicit boundary between background knowledge and observations/results. This normally takes the form of a sentence containing an explicit *Observation* clue word or phrase. The observation interpretation is then normally understood to be "projected" onto events in sentences that follow, that are otherwise unmarked with *Observation* clue words and phrases. The following sentence occurs earlier in the same abstract as (S38):

(S39) We have **found** that ISG *expression* in the monocytic U937 cell line differs from most cell lines previously examined.

The presence of the word *found* in (S39) explicitly indicates that an observation is being described. Sentences that follow but are not explicitly marked with clue words and phrases are highly likely also to describe observations.

- Document Titles

(S40) Leukotriene B4 ***stimulates*** c-fos and c-jun gene transcription and AP-1 binding activity in human monocytes.

Sentence (S40) corresponds to an abstract title. Because of this, it can be assumed that the event centered on the verb *stimulates* is describing new knowledge which has been discovered during the study reported in the paper, and hence the event is assigned the *Observation* category.
NOTE: Events in titles that do not constitute complete sentences (i.e. those without a verb) are generally annotated with the *Investigation* Knowledge Type (see section 2.1.1).

2.1.4 Method

Assigned to events describing experimental methods

- Evidence – Any events whose trigger is a word that describes an experimental method. Typical clue words are *stimulate, stimulation, addition, pretreated* and *incubated* etc. (S41-S42)

NOTES:

- Event triggers DO NOT need to be annotated as clueKT for Method
- Some trigger words (e.g. *stimulate*) do not always happen due to human intervention. Those that occur naturally should not be annotated as *Method*. Often, there is a clue in the event type, i.e. "Artificial_Process", or in the comment attached to the event, which may read "Artificial". Otherwise, if "leads to" or "results in" are used following the "stimulation" event, then this gives a good clue that a method is being described

- Typical position in the text

  Within the section that describes the experiments – normally in the middle section of the paper

- Example Sentences

  (S41) Deoxycholate treatment of the cytoplasmic extract prepared from cells ***stimulated*** by TNF-alpha in the presence of Cu2+ resulted in the release of NF kappa B from I kappa B alpha, indicating that Cu2+ interferes with the dissociation of the NF kappa B-I kappa B complex.

  (S42) In addition, ***pretreatment*** of the cells with the proteasome inhibitor N-Ac-Leu-Leu-norleucinal inhibits this ligand-induced degradation and, in agreement with previous studies, stabilizes a hyperphosphorylated form of the human I kappa B alpha protein.

2.1.5 Fact

Assigned to events that describe general facts and well established knowledge.

- Evidence
  - Events with triggers that describe biological processes in the present tense (unless they describe observations, see section 2.1.3) (S43 – S44)
  - Explicit clue words and phrases are not normally present, with the exception of *known*, which may sometimes be present within the sentence. (S45)

  NOTE: Events of this category can look very similar to those of the *Observation* category (see above). Care should be taken to carefully examine the context of such events before deciding on the most appropriate category to assign.

- Typical Position in the Text

  Normally towards the beginning of the text, describing background knowledge.

- Secondary Events

  Normally *Other*, but may be *Fact* if describing another, complete fact.

- Example Sentences

  (S43) Leukotriene B4 ***stimulates*** c-fos and c-jun gene transcription and AP-1 binding activity in human monocytes.

  (S44) The c-jun mRNA, which is constitutively ***expressed*** in human peripheral-blood monocytes at relatively high levels, was also slightly ***augmented*** by LTB4

  (S45) Oxidants such as hydrogen peroxide are **known** to *activate* certain transcription factors such as nuclear transcription factor kappa beta.

- Discussion of Examples

  When the main event in a sentence or clause corresponds to an observation, *Fact* events can still occur, e.g. to give further factual information which is necessary to fully explain the event. For example, in (S44) the main event of the sentence is centered on *augmented* and is an observation. However, the event centered on *expressed* is providing additional, factual information and so should be annotated as *Fact.*

2.1.6 Other

Assigned to events that do not fit into any other category, those events that do not express complete information, or whose Knowledge Type is unclear or is assignable from the context. Also normally assigned to secondary events, when the Knowledge Type of the primary event is either *Fact* or *Analysis*. The exceptions to this rule are when the secondary event is clearly *Fact* or *Observation.*

- Evidence
  - Secondary events whose primary event has the Knowledge Type of *Analysis, Investigation* or *Fact*. (S46-S47)
  - Secondary events whose primary event has the type *Negative_Regulation* and whose trigger is a word such as *inhibit, prevent, block* or *attenuate,* indicating that the secondary event cannot be said to have taken place. (S48)
  - Secondary events whose primary event is an observation has been negated (i.e., *Polarity = Negative*), but ONLY when this means that the secondary event cannot be said to have taken place (S49). Generally, this rule does not apply when the primary negated observation has the type *Negative_regulation* (S50). In this case, the secondary event can normally be said to be an observation

- Events that describe properties of entities. This is the case in (S51).

- Example Sentences

  (S46) These results **indicate** that LTB4 may *regulate* the production of different cytokines.

  (S47) The *effects* of prostaglandin E2 (PGE2) on cytokine production and proliferation of the CD4+ human helper T cell clone SP-B21 were **investigated**.

  (S48) IL-10 preincubation *inhibited* gene expression for several IFN induced gene

  (S49) Integrin ligation with antibodies does not *induce* tyrosine phosphorylation of FAK.

  (S50) However, no *loss* of DNA binding activity is observed (Not Other)

  (S51) A Rel-related, mitogen-*inducible*, kappa B-binding protein has been cloned as an immediate-early activation gene of human peripheral blood T cells**.**

- Discussion of Examples

  In (S46) the primary event, whose trigger is *regulate*, is an *Analysis* event, according to the presence of the word *indicate*. However, there is a secondary event whose trigger is *production*. The analysis interpretation does not extend to this secondary event, i.e., the interpretation of this event is not that "production of different cytokines *may* occur". In fact, the secondary event does not have a specific interpretation, e.g. there is nothing providing information about whether it is a general fact or under what circumstances it occurs. In other words, it has an incomplete interpretation when considered in isolation from the primary event. For this reason, it is assigned the Knowledge Type of *Other*. Sentence (S47) shows a similar case, where the primary event, whose trigger is *effects*, has the Knowledge Type value of *Investigation*. The secondary events whose triggers are *production* and *proliferation* are thus assigned the type *Other*.

  In (S48), the semantics of the primary event (whose trigger is *inhibit*) mean that the secondary event (with trigger *expression*) did not take place. The same would be true for primary events with triggers *prevent* or *block*.

  In (S49), the fact that primary event (whose trigger is *induce*) is negated, means that the secondary event (with trigger *phospholylation*) did not take place. The primary event is an *Observation* (according to the

context in which it appears). However, the secondary event was not observed, and hence *Other* should be assigned.

In (S50), the primary event, centred on *loss,* describes a negative regulation. The secondary event is centred on *binding activity*. As the primary event is negated, i.e., *no loss of DNA binding activity occurs*, this means that the DNA binding was an observable event, and hence should be annotated as *Observation.*

In (S51), the positive regulation event centered on *inducible* describes a property of the protein, namely that it is induced by mitogen.

2.2 Certainty Level

This dimension aims to identify those events where there is less than 100% certainty that the event will take place (all of the time). This could be for two different reasons:

1) **The author has a lack of (complete) confidence in the truth of the event**. Different levels of confidence can be explicitly specified in the text.
2) **It is believed that the event does not take place all of the time, according to the conditions specified.** In some cases, it is explicitly specified that an event takes place normally or only sometimes, rather than all the time.

Both of the above situations require some kind of cognitive analysis, i.e., the analysis or interpretation of experimental results or other information. It is for this reason that *Certainty Level* values other than the default value can only be assigned to events with a Knowledge Type value *Analysis.*

The default (top level) value of L3, corresponding to complete confidence in the event, is assigned *unless* there are any explicit words or phrases in the sentence that alter the certainty level. That is to say, a certainty level below 100% is *always* expressed using explicit clue words or phrases. Events that are affected in this way are assigned a certainty level of either L2 or L1, depending on the degree of uncertainty expressed.

> **NOTE:** If a sentence contains only a certainty level clue word and not an explicit verb that indicates the Knowledge Type of *Analysis* (e.g. *suggest, indicate,* etc.), then the certainty level clue word should be annotated as **both** a Knowledge Type marker *and* a Certainty Level marker.

The three certainty levels are defined as follows:

2.2.1 L3

The default certainty level category. Assigned to events when there is both:

1) No expression of uncertainty or speculation.
2) No indication that the event does no not occur all of the time (within the conditions/circumstances described).

2.2.2 L2

Assigned to events that either:

1) Express some degree of uncertainty about the truth of the event, but with a confidence level of greater than 50%.
   - Evidence – Always indicated through an explicit word or phrase in same sentence as event. Typical clues are:
     - Words such as *likely* and *probably* (S52-S53).
     - Verbs that are also used as clues for the assignment of the *Analysis* Knowledge Type category, which convey the meaning of a somewhat tentative analysis, e.g. *believe, hypothesize, suggest* and *indicate.*(S54-S55)
2) Express the fact that the event takes place most (but not all) of the time, according to the environmental conditions/circumstances described.
   - Evidence: – Always indicated through an explicit word or phrase in same sentence as event. Typical clues are:
     - Words such as *normally, often, frequently* etc (S56-S57).

- Example Sentences:

(S52) The loss of conventional responsiveness is **probably** *caused* by alterations at the level of signalling

(S53) The MAD-3 cDNA encodes an I kappa B-like protein that is **likely** to be *involved* in regulation of transcriptional responses to NF-kappa B, including adhesion-dependent pathways of monocyte activation.

(S54) Recently, investigators have **hypothesized** that CD14-mediated signaling is *effected* through a receptor-associated tyrosine kinase (TK), suggesting a multicomponent receptor model of LPS signaling.

(S55) During the course of serious bacterial infections, lipopolysaccharide (LPS) is **believed** to *interact* with macrophage receptors, resulting in the generation of inflammatory mediators and systemic symptoms including hemodynamic instability and shock.

(S56) Expression of IL-1alpha by HTLV-I productively infected cells may be important in the hypercalcemia, osteolytic bone lesions, neutrophilia, elevation of C-reactive protein, and fever **frequently** *seen* in patients with HTLV-I-induced adult T-cell leukemia/lymphoma

(S57) HIV-1-infected myeloid cells are **often** *diminished* in their ability to participate in chemotaxis, phagocytosis, and intracellular killing.

2.2.3 L1

Assigned to events that either:

1) Express medium to high uncertainty about the event**,** i.e. the event is interpreted as having a confidence level of 50% or lower.

   - Evidence: – Always indicated through an explicit word or phrase in same sentence as event. Typical clues are:

     - Words such as *may*, *might* and *perhaps* (S58-S59)

       NOTE: If an event is modified by both an explicit *Analysis* verb (e.g. *indicate*), that would by default denote a certainty level of *L2 and* a separate *L1* Certainty Level Marker (e.g. *may),* then the Certainty level value of L1 should be assigned (see S58)

     - Verbs that are also used as clues for the assignment of the *Analysis* Knowledge Type category, which convey the meaning of a highly tentative analysis, e.g., *speculate* (see section 2.1.1).

2) Express the fact that the event takes place only some of the time (normally less that 50%), according to the environmental conditions/circumstances described.

   - Evidence – Always indicated through an explicit word or phrase in same sentence as event. Typical clues are:

     - Words such as *sometimes*, *rarely*, *scarcely*, etc.

- Example Sentences:

  (S58) These results indicate that LTB4 **may** *regulate* the production of different cytokines by modulating the yield and/or the function of transcription factors such as AP-1-binding proto-oncogene products.

  (S59) **Perhaps** murine thymocytes are *denied* this form of rescue because they shut off IL-2R beta chain expression at an earlier stage

2.3 Polarity

This dimension aims to capture whether the event describes a positive or negative situation. We define a negated event as one which describes the absence or non-existence of an entity or a process. That is to say, the event may describe that a process does not or did not happen, or that an entity is absent or does not exist.

There are two possible values of this dimension, namely:

2.3.1 Positive

Where there is no indicated negation of the event (the default category)

2.3.2 Negative

Where the event has been negated, according to the description above.

- Evidence: – Always indicated through an explicit word or phrase in same sentence as event. Typical indicators are:
  - The most common means of expressing negation is through the use of the words *not* or *no* (S61-S62)
  - A number of other words can also be used to express the fact that an event did not take place, when occurring in certain contexts. Examples include *fail*, *lack*, and *unable*, *exception*, *independent*, *without* (S63-S65)
  - **NOTE:** Events that are assigned the type *Negative_Regulation* (centred on verbs such as *inhibit*, *suppress* etc). should NOT be annotated with Polarity=Negative UNLESS there is a specific word or phrase (such as those introduced above) that negates the negative regulation event.

    Although negative regulation events have a negative meaning, this is already encoded in the existing annotation, in the event type (i.e. *Negative Regulation*). As the purpose of the meta-knowledge annotation task is to add information that is not present in the existing annotation, the polarity of negative regulation events should not be annotated as Negative unless the event itself has been explicitly negated.

    As an example, consider (S60), where the marked negative regulation event is centred on the word *inhibits*, but this has not been explicitly negated (i.e. there is no word such as *not* or *no*). In this case, the event should have a Polarity value of *Positive.*

    (S60) Kappa B/MAD-3 completely *inhibits* NF-kappa B p65-dependent transcriptional activation mediated through the human immunodeficiency virus type 1 kappa B enhancer in human T lymphocytes

    In contrast, consider sentence (S61) below. The presence of the word *not* prior to *inhibit* negates the event. In this case, as in the other examples below, the Polarity value should set to *Negative.*

- Example Sentences:

  (S61) CsA was found **not** to *inhibit* lck gene expression, nor the activity of the lck gene product.

  (S62) Protein synthesis inhibitors and corticosteroids, which suppress arachidonate release and the synthesis of proinflammatory cytokines, had **no** *effect* on translocation of NF-kappa B in CHO/CD14 or RAW 264.7 cells, demonstrating that NF-kappa B translocation is an early event.

  (S63) In contrast, NF-kappa B p50 alone **fails** to *stimulate* kappa B-directed transcription, and based on prior in vitro studies, is not directly regulated by I kappa B.

  (S64) The CD19 protein is *expressed* on the surface of all B-lymphoid cells with the **exception** of terminally differentiated plasma cells

(S65) Binding of type I interferon (IFN-alpha/beta) to specific receptors results in the rapid transcriptional activation, ***independent*** of protein synthesis, of IFN-alpha-stimulated genes (ISGs) in human fibroblasts and HeLa and Daudi cell lines.

- Discussion of Examples:

  In sentence (S64), there are 2 events that are centred on the verb *expressed*. In the first event, the CD19 protein is expressed on the surface of all B-lymphoid cells, and so is positive. In the second event, the presence of the word *exception* denotes the fact that CD19 protein is NOT expressed on terminally differentiated plasma cells, and hence should be annotated as a negative event.

  In example (S65), the event centered on the word *independent* has the type *CORRELATION* and involves *transcriptional activation* and *protein synthesis.* The use of the word *independent* itself indicates that *no* correlation exists between them, because the transcriptional activation takes places independently of protein synthesis. Therefore, the correlation event is negative. This example serves to illustrate the potential complexity in recognizing events with negative polarity. Sometimes, the meaning and type of the event have to be considered carefully in order to determine whether it is positive or negative.

2.4 Manner

This dimension aims to identify the rate, level, strength or intensity of the event (in biological terms). We call this the *Manner* of the event, which as three possible values.

NOTES :

1) Manner should only be annotated when it is referring to the rate, level, strength or intensity a biological process.
2) Manner is normally indicated by words other than the event trigger word, unless the meaning of manner is integral to the trigger word, e.g. *overexpression* = expression at a high level.

The words *upregulation* and *downregulation* DO NOT denote high and low manner, respectively. Rather, they denote the direction of the regulation, positive or negative, which is not covered by this annotation dimension, but which is already encoded into the existing event annotation using the types *Positive_Regulation* and *Negative_Regulation.*

2.4.1 High

Assigned to events where there is explicit indication that the event occurs at a high rate, level, strength or intensity.

- Evidence – Always indicated through an explicit word or phrase in same sentence as event, but NOT the event trigger word. Typical clues are:

- Adverbs: examples include *strongly*, *rapidly* and *highly*, etc. (S66-S68)
- Adjectives: examples include *high, rapid, profound*, etc. (S69-S71)

NOTE: If a positive regulation event is triggered by a word such as *enhances*, this trigger word should NOT be annotated as a clue for High Manner, since it denotes only the direction of the regulation (positive rather than negative) and does not say anything about the intensity or level of the regulation. If *enhances* is modified by an adverb like *significantly*, then it is this word that should be annotated as the High Manner marker.

- Example Sentences:

(S66) Both messages **rapidly** *declined* thereafter.

(S67) It was found that lipopolysaccharide *induced* **strongly** both c-fos and c-jun expression.

(S68) Although IFN-gamma alone does not induce ISG expression, IFN-gamma pretreatment **markedly** *increases* and *hastens* ISG expression and transcriptional induction.

(S69) In particular, the c-Rel homodimer *has* a **high** *affinity* for interleukin-6 (IL-6) and beta interferon kappa B sites.

(S70) However, the **profound** T cell *deficit* of nude mice indicates that the thymus is by far the most potent site for inducing the expansion per se.

(S71) Binding of type I interferon (IFN-alpha/beta) to specific receptors results in the **rapid** *transcriptional activation*.

- Discussion of Examples:

Sentence (S65) shows a case where *strongly* indicates a high rate of induction. It is important to remember that *strongly* only indicates a high manner when it is modifying verbs that describe biological processes. When used in conjunction with verbs denoting the *Analysis* Knowledge Type (e.g. *strongly suggest*), it does NOT denote the Manner of the event.

In example sentence (S66), the manner adverb *markedly* applies both to the events centred on *increases* and *hastens*, to indicate a high level.

2.4.2 Low

Assigned to events where there is an explicit indication that the event occurs at a low rate, level, strength or intensity.

- Evidence: – Always indicated through an explicit word or phrase in same sentence as event. Typical clues are:
  - Adverbs: examples include *slightly, partially*. (S70-S71)
  - Adjectives: examples include *little, small, slight*. (S72-S73)
  - Phrases such as *barely, scarcely* (*any*), *almost no*. Although such phrases have negative connotations, they still convey the fact that the stated event took place, even though in a very insignificant way. Hence, the *Polarity* value should be *Positive*, and the *Manner* value should be *Low*. (S74-S75)

- Example sentences

  (S70) The c-jun mRNA was also **slightly** *augmented* by LTB4.

  (S71) Alteration of the sequence at threonine 78 can **partially** *restore function* to a verb A protein rendered defective due to a mutation at position 61.

  (S72) Moreover, kappa 1-kappa 3 can each be deleted from the TNF-alpha promoter with **little** *effect* on the gene's inducibility by PMA.

  (S73) The oxLDL-induced NF-kappa B activation was accompanied by an initial depletion of I kappa B-alpha followed by a **slight** transient *increase* in the level of this inhibitor protein.

  (S74) In contrast, the RelA(p65) subunit was **barely** *detectable* in monocytes, but its level increased markedly in MDMs.

  (S75) Tumor necrosis factor induced slightly c-fos and had **almost no** *effect* on c-jun and AP1.

2.4.3 Neutral

Default category assigned to all events without an explicit indication of manner. However, in rare cases, explicit clues (such as *normal, medium* etc.) could also be found. For example, consider the example sentence (S76).

(S76) The eukaryotic transcription factor NF-kappa B plays a central role in the induced expression of human immunodeficiency virus type 1 and in many aspects of the genetic program mediating **normal** T-cell *activation* and growth.

2.5 Source

This dimension encodes to the source or origin of the knowledge being expressed by the event. Specifically, we wish to distinguish between events that can be attributed to the current study, and those that are attributed to other studies. There are two categories within this dimension, as follows:

2.5.1 Current

The default category. Assigned to events event making an assertion that can be attributed to the current study.

- Evidence
  - Explicit evidence is often not present. Sentences describing results that are unmarked for source normally correspond to *Current*, although this is not exclusively the case, and context must be examined to determine whether the event refers to the current or a previous study.
  - When explicit evidence is present, the word *we* is often present in the sentence. On its own, this is not enough to determine the value of *Current*, as the sentence could be referring to work carried out by the authors in a previous study (see sentence (S80) in the discussion below).
  - Reliable indicators involving *we* include the following:
    - *We have + past_participle*, e.g. *we have found that …* . (S77)
    - The use of *here* in conjunction with *we*, e.g. *we report here that …* denoting that the event is relevant in the current study. (S78)
    - Phrases such as *The present work, in this study,* etc. (S79)

- Example Sentences:

  (S77) **We have** examined the *effect* of leukotriene B4 (LTB4) on the expression of the proto-oncogenes c-jun and c-fos.

  (S78) **We** report **here** that the second alteration, at threonine 78, also *plays* an important, although more indirect, *role.*

  (S79) **The present work** has examined the *effects* of okadaic acid, an inhibitor of type 1 and 2A protein phosphatases, on the regulation of c-jun expression during monocytic differentiation of U-937 leukemia cells.

- Discussion of Examples

  Consider example (S80), which demonstrates how the presence of the word *we* alone is not necessarily sufficient to determine a Source value of *Current:*

  (S80) In addition, **we** looked at the *modulation* of nuclear factors binding specifically to the AP-1 element after LTB4 stimulation.

  In order to determine whether the event marked in (S80) should be annotated as *Current*, the context should be examined. In isolation, the use of the simple past tense (*looked at)* is ambiguous as regards the source, i.e. it may refer to a previous study undertaken by the authors, in which case in would be annotated as *Other* (see below). Equally, it may refer to the current study, in which case it would be annotated as *Current.* However, (S77) and (S80) are drawn from the same abstract, where (S77) immediately precedes (S80). As sentence (S77) contains sufficient evidence to link it to the current study, and as sentence

(S80) is explicitly linked to it through the use of *In addition*, it follows that sentence (S80) must also refer to the current study, and hence should be annotated as *Current.*

Consider an example (S81), where no explicit marker of *Source* is present in the sentence.

(S81) LTB4 ***increased*** the expression of the c-fos gene in a time- and concentration-dependent manner.

Although (S81) is fairly clearly an experimental observation, it is only by examining the context that it can be discovered whether this is a result of the current study, or a previous one. At least for abstracts, if a sentence such as (S77) occurs towards the beginning of the abstract, then it will normally be the case that any subsequently reported results should be interpreted as being attributable to the *Current* study, unless there is any explicit indication to the contrary.

2.5.2 Other

This value indicates that the event is attributed to a previous study.

- Evidence – Always indicated through an explicit word or phrase. Typical clues are:
  - Words and phrases like *previous studies* and *previously* etc. (S82-S83)
  - Citation of another paper (S84)
  - Events that are attributable to the current author, but which implicitly refer to a study other than the current one (S85).
- Example sentences:

(S82) Although it has been **previously** shown that the IL-6 kappa B motif *functions as* a potent IL-1/tumor necrosis factor-*responsive element* in nonlymphoid cells, its activity was found to be repressed in lymphoid cells such as a Jurkat T-cell line.

(S83) Since **previous studies** have demonstrated that the c-jun gene is *autoinduced* by Jun/AP-1, we also studied transcription of c-jun promoter (positions -132/+170)-reporter gene constructs with and without a mutated AP-1 element.

(S84) A recent functional analysis by **Miyatake et al. (S. Miyatake, M. Seiki, M. Yoshida, and K. Arai, Mol. Cell. Biol. 8:5581-5587, 1988)** described a short promoter region in the GM-CSF gene that conferred strong *inducibility* by T-cell-activating signals and tax1, but no NF-kappa B-binding motifs were identified.

(S85) We have **earlier** found that in Jurkat cells activation of protein kinase C (PKC) *enhances* the cyclic adenosine monophosphate (cAMP) accumulation induced by adenosine receptor stimulation or activation of Gs.

In (S85), although the use of the present perfect "we have" would normally indicate that the reported event belongs to the current study, the presence of the word *earlier* shows that event centred on *enhances* is an observation from an earlier study.

3 Hypothetical Examples

Having examined in the different annotation dimensions of the scheme in some detail, we now re-examine the hypothetical sentences first introduced in section 1.2.1, and discuss the correct categories to assign to them for each meta-knowledge dimension.

(S3) *We **found** that Y activates the expression of X*

Event 1: *activates*

Knowledge Type: ***Observation***. The word *found* shows that the event corresponds to an observed result.

Certainty Level: ***L3***. There are no words or phrases to suggest that the event does not take place all of the time, and so the default value of *L3* is assigned

Polarity: ***Positive.*** There are no words or phrases expressing the negation of the event, so the default value of *Positive* is assigned.

Manner: ***Neutral***. There are no words or phrases expressing manner, hence the default value of *Neutral* is assigned

Source: ***Current.*** In isolation, this sentence is ambiguous between a source value of *Current* or *Other*. However, in this and other examples in this section that are unmarked as regards their source, we assume that the context allows the value of *Current* to be assigned.

Event 2: *expression*

Knowledge Type: ***Observation.*** Inherited from Event 1. If a top-level event is assigned the *Observation* category, then its sub events will also normally be assigned this category, unless there is any other evidence in the context to suggest otherwise.

Certainty Level: ***L3***.

Polarity: ***Positive.***

Manner: ***Neutral***.

Source: ***Current.***

(S4) *We **examined** the effect of Y on expression of X*

Event 1: *activates*

Knowledge Type: ***Investigation***. The word *examined* shows that the event corresponds to something that is to be investigated

Certainty Level: ***L3***. This dimension is not applicable to *Investigation* events, and so the default value is automatically assigned

Polarity: ***Positive.***

Manner: ***Neutral***. This dimension is not applicable to *Investigation* events, and so the default value is automatically assigned

Source: ***Current.***

Event 2: *expression*

Knowledge Type: ***Other***. This event does not directly correspond to what is being investigated. Nothing is being said about the intended interpretation of this event, and so ***Other*** is assigned

Certainty Level: ***L3***.

Polarity: ***Positive.***

Manner: ***Neutral***.

Source: ***Current.***

(S5) *These results **suggest** that Y has **no** effect on expression of X*

Event 1: *effect*

Knowledge Type: ***Analysis***. The word *suggest* with the subject *These results* shows that the event corresponds to an analysis of the results.

Certainty Level: ***L2***. This word *suggest* shows that the analysis that has been made is somewhat tentative, and so *L2* is assigned.

Polarity: ***Negative.*** The presence of the word *no* before the event trigger word negates the event.

Manner: ***Neutral***.

Source: ***Current.***

Event 2: *expression*

Knowledge Type: ***Other***. The analysis interpretation only applies to event 1. Nothing is being said about the intended interpretation of this event. As the top level event is *Analysis*, it cannot be considered as a fact, nor is it an observation. Hence, ***Other*** is the most appropriate category to assign.

Certainty Level: ***L3***.

Polarity: ***Positive.***

Manner: ***Neutral***.

Source: ***Current.***

(S6) *Y is **known** to increase expression of X*

Event 1: *increase*

Knowledge Type: ***Fact***. The presence of the word *known* makes explicit that event corresponds to a generally accepted fact.

Certainty Level: ***L3***. .

Polarity: ***Positive***

Manner: ***Neutral***.

Source: ***Curren*t.**

Event 2: *expression*

Knowledge Type: ***Other***. The factual interpretation only applies to event 1. Nothing is being said about the intended interpretation of this event, and so ***Other*** is assigned

Certainty Level: ***L3***.

Polarity: ***Positive.***

Manner: ***Neutral***.

Source: ***Curren*t.**

(S7) *Addition of Y **slightly** increased the expression of X*

Event 1: *increased*

Knowledge Type: ***Observation***. The use of the past tense on the trigger word signifies that this is an experimental observation

Certainty Level: ***L3***. .

Polarity: ***Positive***

Manner: ***Low***. The use of the word *slightly* indicates the amount of increase is small, and so the value of *Low* is assigned

Source: ***Curren*t.**

Event 2: *expression*

Knowledge Type: ***Observation***. As event 1 is an observation, so is event 2.

Certainty Level: ***L3***.

Polarity: ***Positive.***

Manner: ***Neutral***.

Source: ***Curren*t.**

(S8) *These results **suggest** that Y **might** affect the expression of X*

Event 1: *affect*

Knowledge Type: ***Analysis***. The use of the verb *suggest* with the subject *These result* marks this event as an analysis

Certainty Level: ***L1***. Although the default certainty level for *suggest* is *L2*, the presence of the word *might* lowers the certainty level to L1

Polarity: ***Positive***

Manner: ***Neutral***.

Source: ***Curren*t.**

Event 2: *expression*

Knowledge Type: ***Other***. Nothing specific is said regarding the interpretation of this event.

Certainty Level: ***L3***.

Polarity: ***Positive.***

Manner: ***Neutral***.

Source: ***Curren*t.**

(S9) ***Significant*** *expression of X was* ***observed***

Event 1: *expression*

Knowledge Type: ***Observation***. Clearly indicated through the use of the verb *observed*

Certainty Level: ***L3***.

Polarity: ***Positive***

Manner: ***High***. The presence of the word *significant* shows that the rate of expression is higher than normal.

Source: ***Curren*t.**

(S10) ***Previous studies*** *have* ***shown*** *that Y activates the expression of X*

Event 1: *activates*

Knowledge Type: ***Analysis***. The verb *shown* is present, with the subject of *Previous studies.* As this is an inanimate subject, the intended interpretation is that some analysis has been undertaken in order to be able to state the event

Certainty Level: ***L3***. Although some analysis clue words convey an L2 certainty level, the verb *shown* does not convey any uncertainty in the analysis, and so a certainty level value of L2 is assigned.

Polarity: ***Positive***

Manner: ***Neutral***.

Source: ***Other*.** The use of the phrase *Previous studies* explicitly shows that the event is attributable to another study.

Event 2: *expression*
Knowledge Type: ***Other***.

Certainty Level: ***L3***.

Polarity: ***Positive.***

Manner: ***Neutral***.

Source: ***Current.*** Although event 1 has a source value of *Other*, here we leave the default value of *Current*, as nothing is being said specifically about the interpretation of this event.

Figure 3 shows all of the above sentences and their respective annotations

(S3) *We **found** that Y activates the expression of X*

(S4) *We **examined** the effect of Y on expression of X*

(S5) *These results **suggest** that Y has **no** effect on expression of X*

(S6) *Y is **known** to increase expression of X*

(S7) *Addition of Y **slightly** increased the expression of X*

(S8) *These results **suggest** that Y **might** affect the expression of X*

(S9) ***Significant** expression of X was **observed***

(S10) ***Previous studies** have **shown** that Y activates the expression of X*

Sentence ID	E2					E1				
	Knowledge Type	Certainty Level	Polarity	Manner	Source	Knowledge Type	Certainty Level	Polarity	Manner	Source
S3	Observation	L3	Positive	Neutral	Current	Observation	L3	Positive	Neutral	Current
S4	Investigation	L3	Positive	Neutral	Current	Gen-Other	L3	Positive	Neutral	Current
S5	Analysis	L2	Negative	Neutral	Current	Gen-Other	L3	Positive	Neutral	Current
S6	Gen-Fact	L3	Positive	Neutral	Current	Gen-Other	L3	Positive	Neutral	Current
S7	Observation	L3	Positive	Low	Current	Observation	L3	Positive	Neutral	Current
S8	Analysis	L1	Positive	Neutral	Current	Gen-Other	L3	Positive	Neutral	Current
S9	Observation	L3	Positive	High	Current	-	-	-	-	-
S10	Analysis	L3	Positive	Neutral	Other	Gen-Other	L3	Positive	Neutral	Current

Figure 3 – Hypothetical Sentences and their Annotation

4 Annotation Task

In the previous section, the annotation was annotated from a slightly abstract point of view, in that detailed information was not given regarding the events on top of which the meta-knowledge will be performed, or about the steps involved in the annotation task. This section addresses these aspects of the task in more detail.

4.1 What Annotation is Already There?

The annotation of meta-knowledge will be performed on a corpus consisting of MEDLINE biomedical abstracts that have already been annotated with events by domain experts. This corpus is called the GENIA event corpus. The event annotation of GENIA was carried out using an annotation tool called X-Conc. This same tool will be used to carry out the meta-knowledge annotation. Below, we provide further information regarding GENIA events. To help with this, Figure 4 illustrates a screenshot of events that have been annotated using X-Conc.

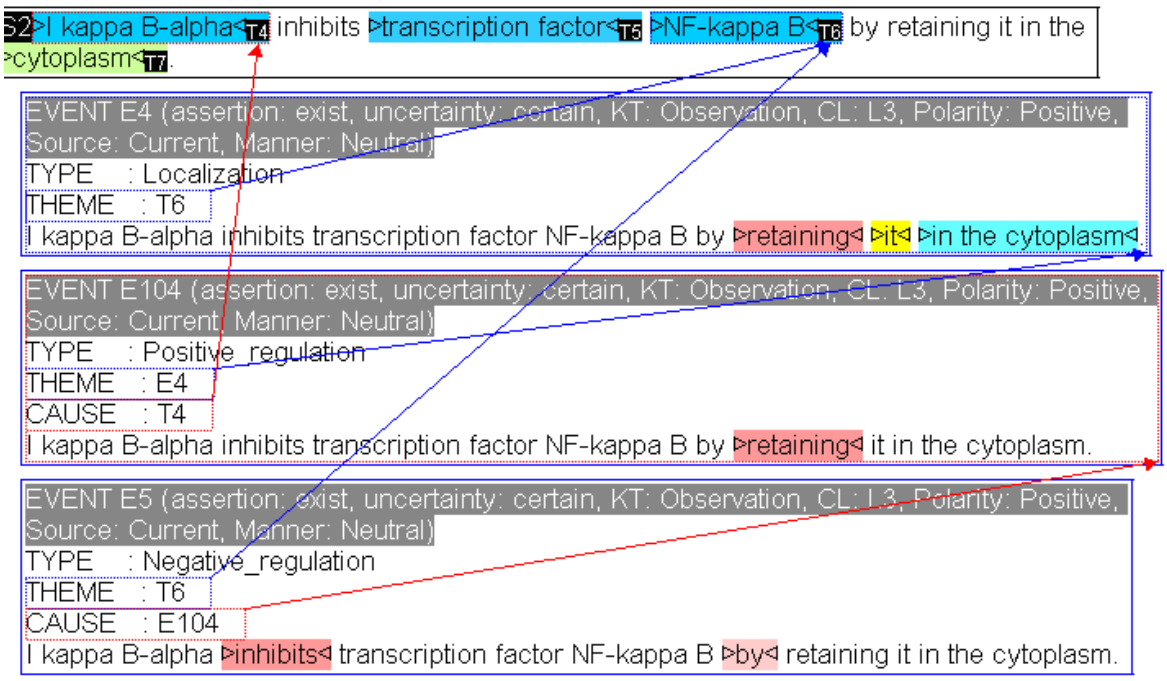

Figure 4 – **Annotated** events in X-Conc

4.1.1 Named Entity Annotations

Each sentence in the abstract is displayed in a box, together with an alphanumeric identifier (id) (e.g. *S2* in the example shown in Figure 4). Within this box, named entities (NEs) are highlighted. In the example shown, entities with blue backgrounds correspond to proteins, and those with green backgrounds correspond to cellular locations. Several other background colours may be used according to different categories of NEs. These entities may correspond to participants in events. Each NE is itself assigned an alphanumeric id. For example, *I kappa B-alpha* is assigned the identifier *T4*.

4.1.2 Event Annotations

Below the box containing the sentence and its NE annotations are boxes corresponding to event annotations (i.e., event frames). Each box repeats the text of the sentence. Highlighted words and phrases correspond to text-span annotations added as part of the original event annotation process; these are described below. Each event is assigned an alphanumeric id (e.g. *E4).* In each event frame box, there are 4 principal types/zones of information

1) ***Event Interpretation/Meta-Knowledge Annotation:*** comes at the top of each event frame box with a grey background colour (also showing the event ID). Shows a set of attribute value pairs. Whilst the majority of these correspond to meta-knowledge annotation dimensions, the remaining two attributes were added as part of the original GENIA event annotation, providing rudimentary information regarding the interpretation of the event.

    a) *Assertion:* Has 2 possible values: *exist* (for positive events) and *non-exist* (for negative events). This is somewhat similar to our *Polarity* dimension, although the values will not always be the same, due to different definitions, and our more fine-grained annotation scheme. For example, some events annotated as *non-exist* in the original GENIA annotation will, according to our meta-knowledge annotation scheme, have a *Polarity* value of *Positive* and a *Manner* value of *Low*. Thus, it SHOULD NOT be assumed that *Polarity* should be set to *Negative* whenever *assertion* is set to *non-exist.* The context of the event should be carefully studied and the guidelines followed in order to assign the correct value for *Polarity*.

    b) *Uncertainty:* Has 3 possible values: *certain*, *probable* and *doubtful.* Somewhat similar to our *Certainty Level* dimension, but again with different values and different definitions. For example, most events annotated as *doubtful* correspond to events that would be assigned a *Knowledge Type* of *Investigation* in our scheme. *Probable* events, meanwhile, could correspond to *L1* or *L2* events in our scheme, if explicit markers are present. Certainty level expressed through analysis markers like *suggest* is not covered by the existing GENIA scheme

        Note: Although these attributes have some aspects in common with some of our annotation dimensions, the fact that our scheme is different means that there are not always direct correspondences, as explained above. For this reason, it is recommended to ignore these when performing meta-knowledge annotation.

2) ***Type*:** This is shown immediately below the interpretation/meta-knowledge. It corresponds to a value assigned from the GENIA event ontology, which is a hierarchical structure of 36 different event types, as shown in Figure 5

3) ***Event Participants*:** Generally these correspond to the THEME and/or CAUSE roles. In X-Conc, each participant role is shown together with the id of the participant, either an NE or another event. Arrows also link the IDs to their actual occurrences, i.e., either to the highlighted NEs displayed in the sentence box above the events, or one of the other event frames.

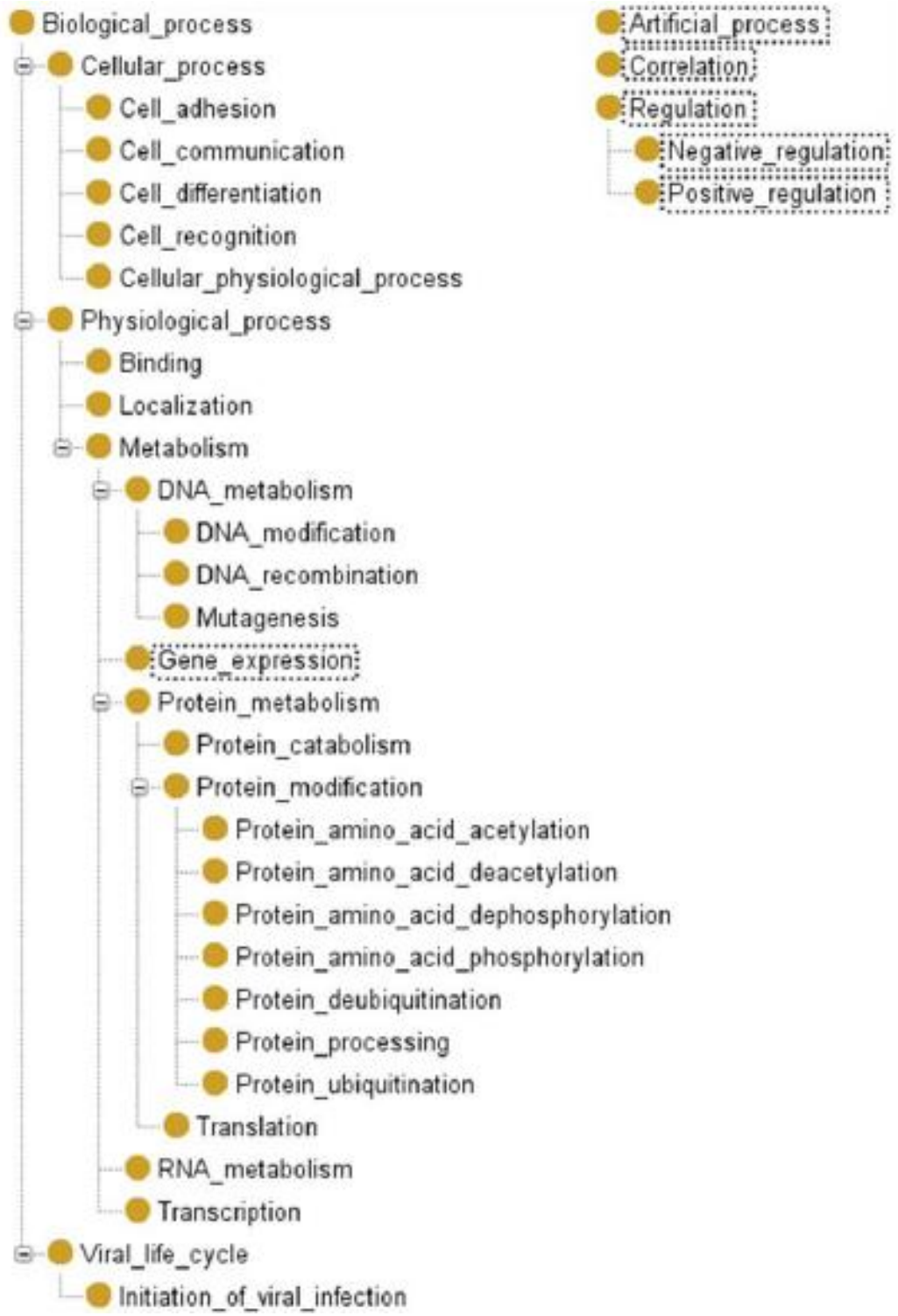

Figure 5 – GENIA Event ontology

4) ***Clue*:** This consists of the complete sentence with text span annotations corresponding to various types of information:

   a) *clueType* – The event trigger word or phrase. This is the word or phrase around which the event is organised, or which can be said to characterise the event. This is also always present and is shown with a dark pink background.

b) *clueLoc* – corresponds to the location in which the event took place. Shown with a cyan background colour.

c) *clueExperiment* – corresponds to experimental techniques specified for the event. Shown using a peppermint green background colour.

d) *clueTime* – corresponds to when the event happened or will happen. Shown using a violet background colour.

e) *linkCause* – used to indicate words that are used in the text link between and event and its CAUSE. They can be seen as words that "introduce" the CAUSE of the event, Typical examples include the prepositions *by, through, with.* Shown using a pink/purple background. Example: *Activation of NFkB **by** IL-2*

f) *linkTheme* – used to indicate words used in the text to link the event and its THEME. They can be seen as words that introduce the THEME of the event. Typical examples include the prepositions *of, in* and *on.* Example: *transcription **of** NFkB.* Shown using a cream background.

g) *coRefCause* – annotated when the CAUSE of the event is an expression such as *it* or *this protein*, referring to a previously introduced (or *coreferent*) NE, either in the current sentence or in a previous sentence. The id specified for the CAUSE role is the id of the original mention of the NE, whilst the co-referring expression will be highlighted in the text using a purple colour.

h) *coRefTheme* – annotated when the THEME of the event contains an expression such as *it* or *this protein*, referring to a previously introduced (or *coreferent*) NE, either in the current sentence or in a previous sentence. The id specified for the THEME role is the id of the original mention of the NE, whilst the co-referring expression will be highlighted in yellow

Having described the main features of the GENIA event representation, we will now describe in more detail the example events shown in figure 4. For ease of explanation, the sentence is repeated below:

"I kappa B-alpha inhibits transcription factor NF-kappa B by retaining it in the cytoplasm"

E4

This is an event assigned the type *Localization*. This type of event provides information regarding the location of a protein. The THEME corresponds to the entity whose location is being described. This THEME is the entity with id *T6*, which is *NF-Kappa B.* Note that the THEME icorresponds to the specific entity name, rather than the more general description, i.e. *transcription factor*. In the *clue* element, i.e., the sentence text, we can see that three text spans have been highlighted in different colours. These are as follows:

- The verb *retaining* corresponds to the *clueType* (i.e. event trigger). This is the verb that is most closely associated with the description of the location (dark pink)

- The word *it* has been annotated as the *coRefTheme* (yellow). This is because *it* as acting as the THEME of the event (since it is the grammatical object of the verb *retaining*). However, *it* itself is not an NE, but rather refers to the previously mentioned NE *NF-Kappa B.* Therefore *NF-Kappa B* is the actual THEME of the event, but this THEME is linked though annotation
- The phrase *in the cytoplasm* has been annotated as *clueLoc* (cyan background). This provides the location information for the *NF-Kappa B* protein.

E104

This event is assigned the type *Positive_regulation.* It has the same *clueType* as E4, i.e. *retaining*, but the actual event is different and with different participants, with a CAUSE as well as a THEME. The CAUSE is the NE with the id *T4*, which corresponds to *I kappa B-alpha*. CAUSEs often correspond to the grammatical subjects of verbs, but meaning as well as grammar is considered during annotation. So, although *I kappa B-alpha* is grammatically the subject of *inhibits*, it can also be seen as the subject of the verb *retaining* when meaning is taken into account. That is to say, one of the facts that can be understood from reading the sentence is: *I kappa B-alpha retains NF-kappa B in the cytoplasm.* It is this fact that corresponds to the event E104. Therefore *I kappa B-alpha* is the CAUSE, whilst the THEME is E4, which is the previously annotated *Localization* event.

E5

This event is assigned the type *Negative_regulation*, based on the *clueType inhibits.* The THEME (the thing being inhibited) is the entity *NF-kappa B* (id *T6).* Although *I kappa B-alpha* is grammatically the subject of *inhibits*, and so could be seen as the CAUSE of the event, *E104* corresponds to the complete event describing *how* this inhibition occurs. Therefore, it is E104 that is annotated as the CAUSE of E5. Within the *clue* element, the word *by* is annotated as *linkCause*, because the preposition that introduces the *clueType* of the event that forms the CAUSE of the event, i.e., *retaining.*

4.2 What to Annotate

The annotation of meta-knowledge should be carried out for *every* event that has been annotated in the document. The task consists of two parts:

1) Assignment of an appropriate value for each meta-knowledge dimension.
2) Annotation of clue words and phrases (if any) that give evidence for the assignment of the appropriate meta-knowledge value.

Section 2 explained in some detail the possible values of each meta-knowledge dimension. Section 5 will explain the practicalities of carrying out these tasks within the X-Conc annotation tool. Below, we provide more detailed information about the suggested sequence of annotation, together with a more detailed explanation of what constitutes a clue phrase.

4.2.1 Sequence of annotation

Within each sentence, there are usually two types of events

1) Primary events, which describe the main assertions in the sentence. These events normally describe well-established knowledge (KT=*Fact*), observations, analyses of results (KT=*Analysis*) or investigations. Such events are normally (but not always) triggered by verbs
2) Secondary events, which form participants of the primary events. Whilst primary can be seen as constituting "complete" facts or assertions, secondary events normally provide only partial information, which can only be correctly interpreted in the context of the primary event. Such events are often (but not always) triggered by nominalised verbs.

Consider the following sentence:

LTB4 increased the expression of the c-fos gene in a time- and concentration-dependent manner.

There are 2 events in this sentence, one with the trigger *increased,* and one with the trigger *expression.* The events have the following structure:

ID: E1
Type: POSITIVE_REGULATION
Trigger: *increased*
CAUSE: *LTB4*
THEME: E2

ID: E2
Type: GENE_EXPRESSION
Trigger: *expression*
THEME: *c-fos gene*

The event E1 is the primary event here, as it constitutes the main observation described in the sentence. E2 is a *secondary* event, as it is a *participant* event of E1 and taken in isolation, does not express complete information. It only makes sense when combined with E1.

Rather than annotating events sequentially as they appear in a sentence, it is suggested that the best way to annotate is the following:

1) Examine all events in a sentence, and locate firstly those that correspond to *primary* events. This is because the KT value assigned to the primary event frequently determines the KT value assigned to the secondary event.
    a) Examine carefully the participants of each event, i.e. the THEME (and CAUSE, if present), as well as the event type (e.g. POSITIVE_REGULATION).
    b) Try to understand the information being conveyed by the event describing a complete fact or assertion. If so, it should be treated as a primary event. If not, then it is a secondary event.

        NOTE: In some sentences, there may be no primary event annotated. In this case, the KT of *Other* should be assigned to the secondary events in

the sentence, unless the textual context of the event strongly suggests that these events should annotated with one of the other KT values.

c) Participant events of primary events are secondary events, and the KT of these secondary events should be assigned according to the rules described below

NOTES

i) Primary events will normally NOT form participants of other events.
ii) The KT value of primary events will always be *Fact*, *Observation*, *Analysis* or *Investigation*.

3) According to the KT value assigned to a primary event, specific KT values will normally be assigned to secondary events that form participants of the primary event. These values are as follows:

a. Primary event: KT=Observation, secondary event: KT =Observation

Example:
RFLAT-1 ***activates*** *RANTES gene expression in T lymphocytes.*

In the above sentence, both the positive regulation event triggered *activates* and the secondary event triggered by *expression* can be said to have been observed.

EXCEPTION: If the context determines that the secondary event DID NOT take place, then the secondary event (normally the THEME) should have KT= *Other*.

Examples:

1) Certain trigger words for the primary event determine that the THEME of the primary event (when this is event) did not take place. Examples of such trigger words for primary events include: *inhibit, prevent, block,* e.g.
IL-10 preincubation ***inhibited*** *gene expression for several IFN-induced genes*

In the above example, the fact that the primary event is triggered by *inhibted* means that the secondary *expression* event cannot be said to have taken place.

2) Certain instances when the primary event is negated mean that the secondary event is not observable, e.g.:

NF-kappa B p50 alone fails to ***stimulate*** *kappa B-directed transcription*

In the above example, the primary event, *stimulate,* is negated by the phase *fails to.* This means that the stimulation did not happen, and hence the secondary transcription event cannot be said to have been observed.

BEWARE: A negated primary event does not always mean that its THEME should be annotated as *Other*. The meaning must be carefully considered, e.g..

IL-10 preincubation ***did not inhibit*** *gene expression for several IFN-induced genes*

In the above example, the fact the primary event has been negated means that inhibition did not occur, and hence the gene expression event CAN be annotated with KT=*Observation*.

b. Primary event: KT=Analysis, Investigation, Fact, Secondary event: KT= Other (**except** if the secondary event describes a method, in which can Method can be used)

If the primary event has a KT value of anything other than *Observation*, then the KT of participant secondary event should normally be *Other* or *Method*, UNLESS there is strong contextual evidence that a different KT value should be assigned.

These rules are summarised in the table in section 6.

4.2.2 Annotating Clue Phrases

If the value of a particular annotation dimension has been assigned on the basis of a word or phrase in the same sentence as the event, then this word or phrase should be annotated as such. As part of the annotation process, clue phrases that are annotated are categorized according to the dimension for which they provide a clue, i.e. *clueKT* (for Knowledge Type), *clueCL* (for Certainty Level), *clueManner*, *cluePolarity* and *clueSource*. In this section, we clarify the types of words and phrases which should be annotated as clues, and set down some rules about the exact text spans to be annotated.

1) There may be several types of evidence which can be used to determine the value of a particular dimension. Only the most "reliable" evidence should be annotated. There are two types of evidence that have been identified for the assignment of a particular value to a dimension:
   a) Explicit clue words or phrases
   b) The event trigger word(s) (e.g., verbs in the past tense which describe biological processes most often denote a *Knowledge Type* of *Observation*), or verbs that denote some kind of Analysis.

   Explicit words or phrases are generally more reliable evidence than event trigger words. Therefore, event trigger words should only be annotated as clue phrases if no other explicit evidence is present.
2) Clues are NOT to be annotated for the default categories along each dimension. Table 1 shows the default categories for each dimension:

Dimension	Default Category
Knowledge Type	Other
Certainty Level	L3
Polarity	Positive
Manner	Neutral
Source	Current

Table 1 – List of Default Categories for Each Dimension

3) As a general rule, the contents of the "clue" annotation should be the minimum unit of text which can be used to determine the correct value for the given annotation dimension.

4) The *clueType* (i.e. the event-trigger) itself should *only* be annotated as a meta-knowledge clue if it "explicitly" represents a meta-knowledge category. So, if the *Observation* Knowledge Type category is assigned on the basis of the *clueType* verb being in the past tense, then this event *clueType* should NOT be annotated as a meta-knowledge clue.

5) Where possible, a single word should be annotated as the clue phrase (e.g., the value of the *Manner* dimension is normally indicated through adverbs or adjectives, whilst the value of the *Knowledge Type* attribute is often indicated through the use of a verb)

6) If the clue phrase is a phrasal verb (e.g. *looked at*), then both the verb and its following preposition should be annotated as the clue phrase.

7) If the clue phase is part of a group of verbs, e.g. *have examined,* then it is only necessary the actual verb which helps to determine the dimension value, unless the tense indicated by the verb group has a bearing on the value of the attribute. Consider sentence (S78), where we are concentrating on the event centered on "effect":

(S86) **Previous studies** have **examined** the *effect* of leukotriene B4 (LTB4) on the expression of the proto-oncogenes c-jun and c-fos.

Within the sentence, the context of the event can help us to determine the values of both the *Knowledge Type* and the *Source* attributes:

a) The *Knowledge Type* dimension of *Investigation* would be assigned whether the past or the present perfect tense were used, i.e. whether the sentence begins *we examined* or *we have examined.* Therefore, for the *Knowledge Type* attribute, only the word *examined* needs to be annotated.

b) For the *Source* dimension, it is the noun phrase *previous studies* that allows us to determine that the event is attributable to some other source (i.e. the assignment of the *Other* category*)*. Therefore, the clue span for the *Source* dimension should consist of the entire noun phrase *previous studies.*

5 Annotation Environment

5.1 Introduction to X-Conc

XConc Suite is a collection of tools supporting the manual annotation a corpus. It runs as a "plug-in" inside the Eclipse application, which is a software development environment.

5.1.1 Getting Started

In order to annotate documents, you will need a copy of the Eclipse application. A copy of Eclipse including the X-Conc plug in will be provided to you. If you have your own version of Eclipse, you can install the X-Conc plugin using the following steps.

1) From the main menu, select **Help** > **Software Updates** > **Find and Install...**.
2) Select **Search for new features to install**.
3) Select **New Remote Site...** and enter a name and URL http://www-tsujii.is.s.u-tokyo.ac.jp/xconc/ .
4) Check the name and click **Finish**.
5) Select XConc Suite, agree the licence, and click **Finish**.
6) The XConc Suite will be installed after re-launching Eclipse.

To start Eclipse, go to the "eclipse" directory and double click on the "eclipse.exe" icon. When you start Eclipse, you will be prompted to enter a "workspace" directory, as shown in Figure 6

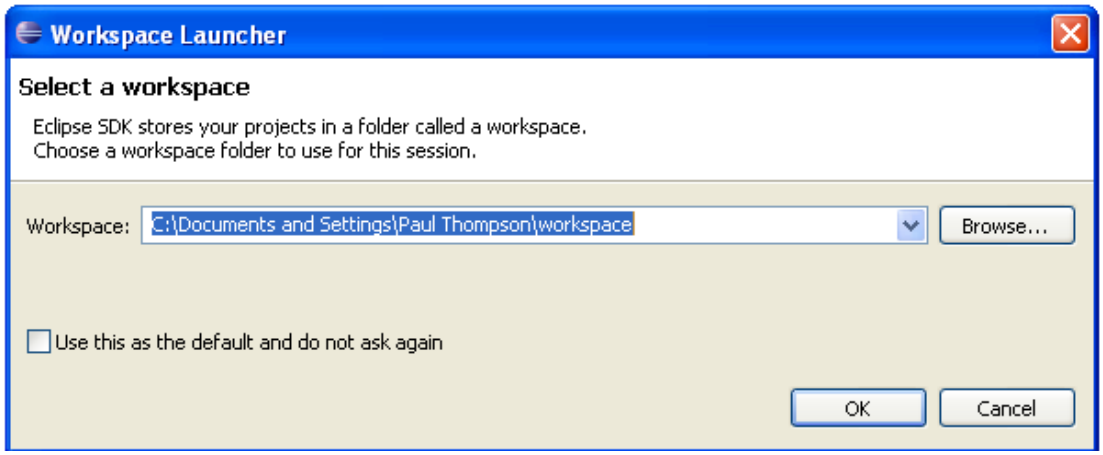

Figure 6 – Workspace Launcher Window

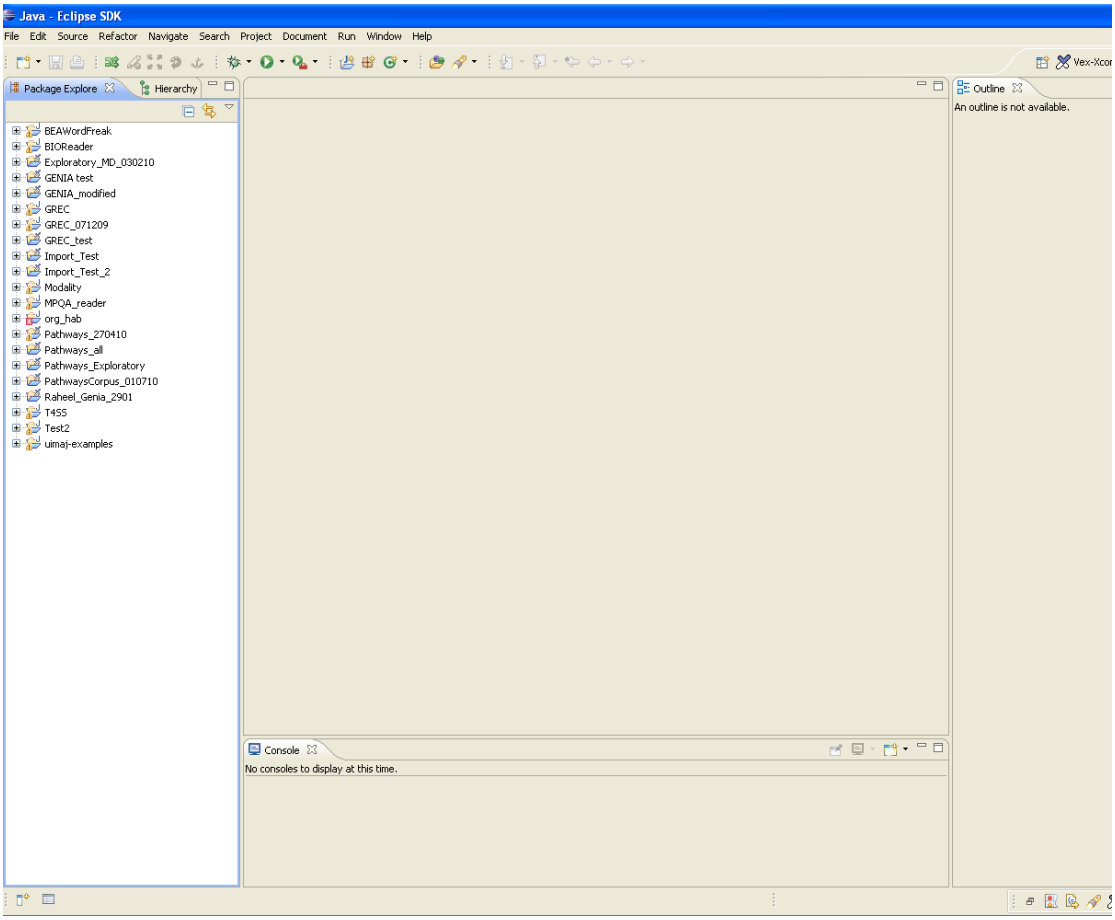

Figure 7 – Eclipse Main Window

5.1.2 Importing Annotation Projects

To carry out annotation, you need to import or create a project. We will provide you with projects to import. You should carry out the following steps:

1) Unzip the project folder to a location of your choice on your computer
2) Choose the "Import .. " option from the "File" menu.
3) From the "Import" window that appears (Figure 8), select the item "Existing Projects into Workspace" under "General"
4) In the "Import Projects" window that appears (Figure 9), ensure that the "Select root directory" option has been selected
5) Click on the "Browse" button next to "Select root directory". A "Browse for folder" window will appear. Browse to the directory where the project has been unzipped. Select this directory, and then click on "OK" at the bottom of the window.
6) In the "Import projects" window, and item corresponding to the selected project should appear in the "Projects" box.
7) Check the box labelled "Copy projects into workspace", and then click on "Finish"

8) The name of the project should them appear in the “Package Explore” on the left-hand side of the man Eclipse screen

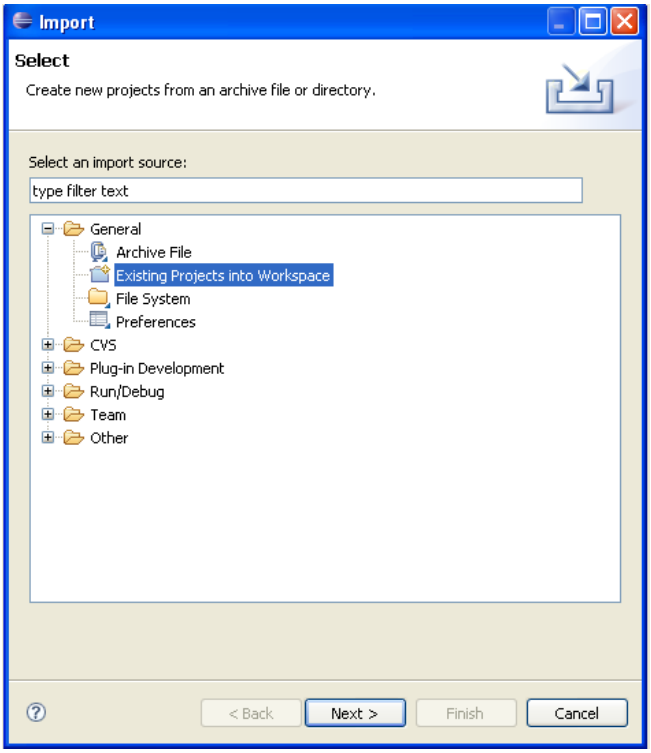

Figure 8 – Import Window

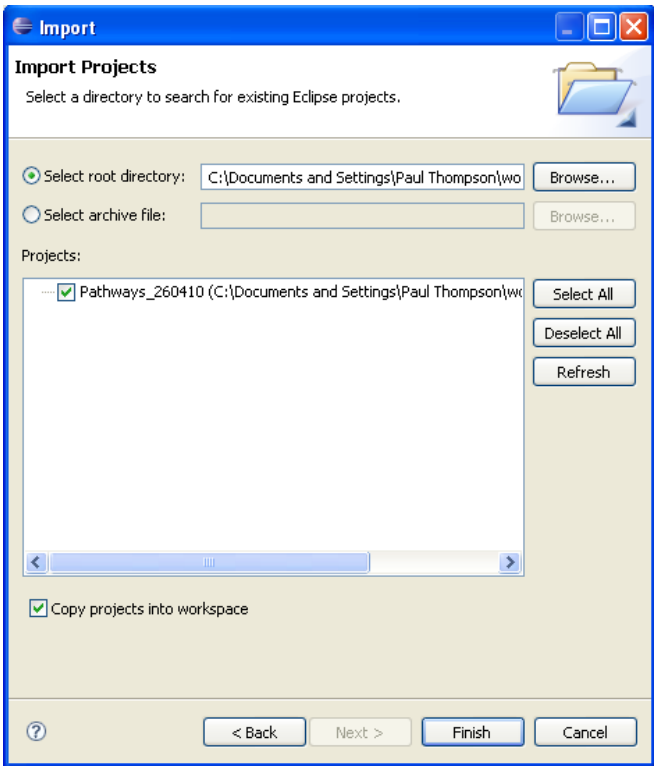

Figure 9 – Importing

5.1.3 Getting Ready to Annotate

Expand the imported project in the "Package Explore" window by clicking on the "+" sign next to the name. There should be 2 folders, one called "Corpus" and the other called "ModifiedGENIATypes", along with a file called "veg-plugin.xml". Expand the "Corpus" folder to see the names of the files to annotate. DO NOT edit the files inside the "ModifiedGENIATypes" folder. These control the display of the annotations.

To begin to annotate, double click on one of the file names within the "Corpus folder". The file should be displayed in a graphical format (as in Figure 11). If the file contents are displayed as text only, then right click on the filename, and select the "Open with" option, and then choose "Vex XML editor". This is illustrated in Figure 10.

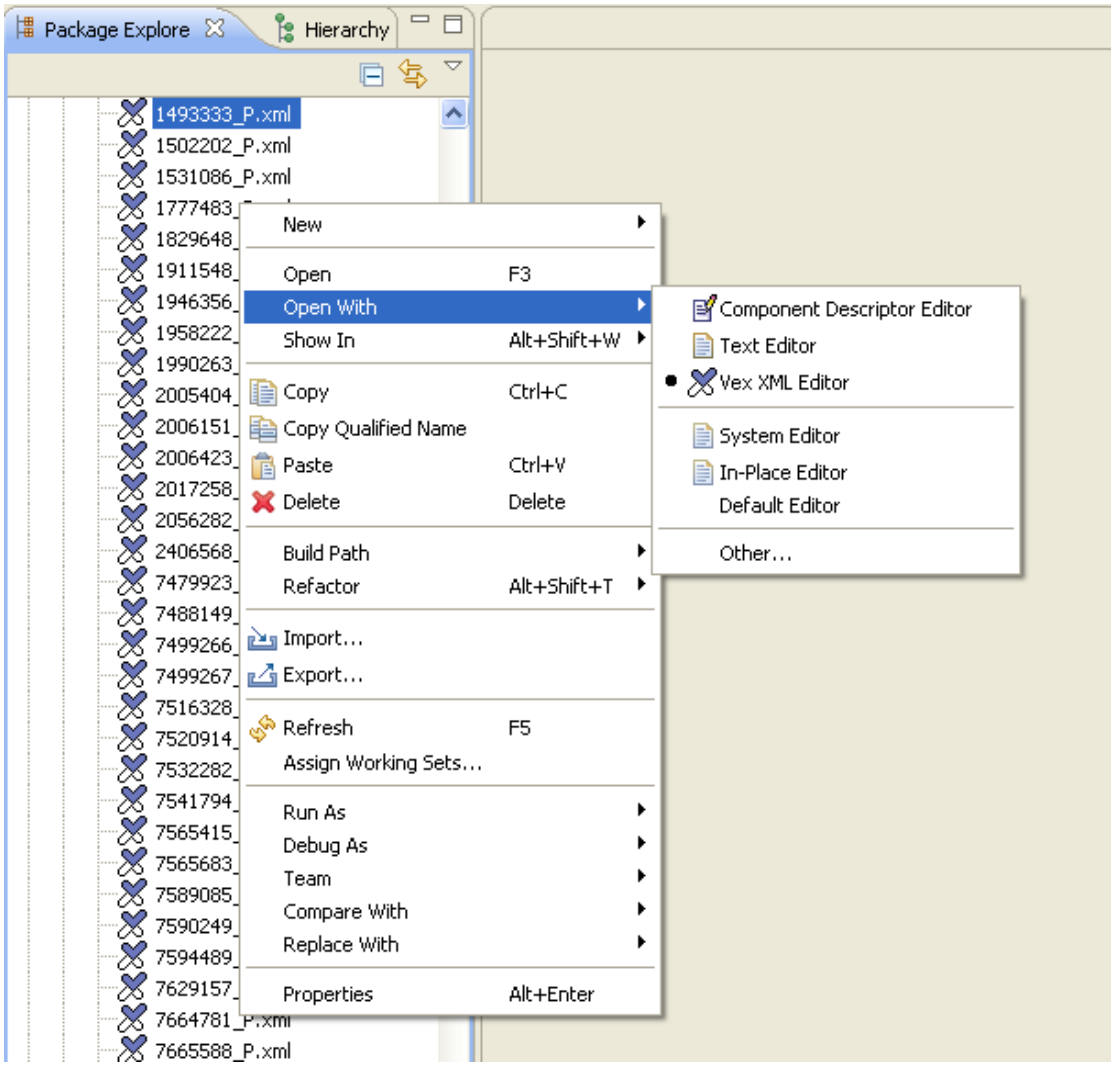

Figure 10 – Choosing to view a file with the Vex XML editor

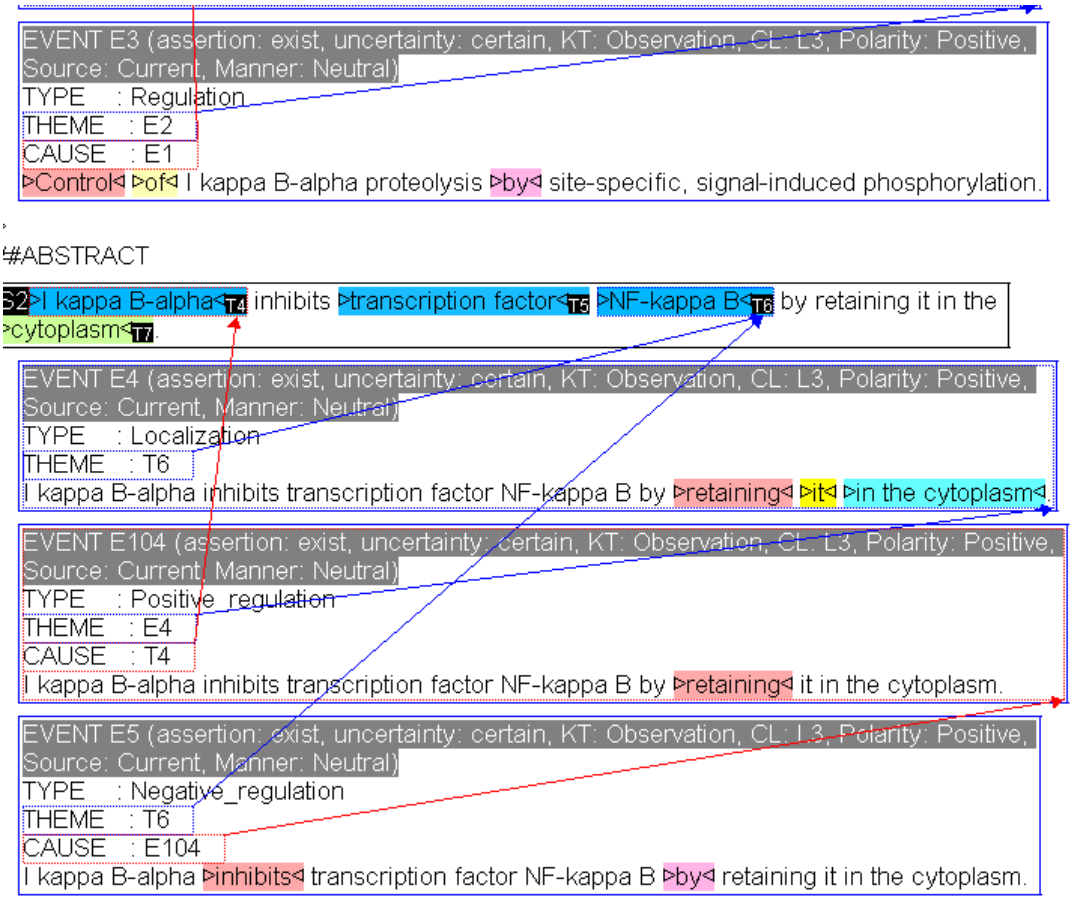

Figure 11 – The Vex view of Document Annotation

5.2.2 Annotating Meta-Knowledge Dimension Values

The grey section at the top of each event (see Figure 11) shows the currently assigned values for the 5 meta-knowledge dimensions, in addition to the 2 event interpretation attributes, i.e. *assertion* and *uncertainty* (these were explained in section 4.1.2, but should NOT be used to influence decisions made during the meta-knowledge annotation process).

5.2.3 Editing Meta-Knowledge Dimension Values

Each of the 5 meta-knowledge dimensions is automatically assigned a "default" in each event. The default value generally corresponds to the most common value for the dimension. The default values are as follows:

KT: *Observation*

CL: *L3*

Polarity: *Positive*

Source: *Current*

Manner: *Neutral*

If any of these values need to be edited, then the following steps need to be taken:

1) Click with the mouse anywhere inside the grey area at the top of the box containing the event representation.

2) Ensure that the cursor is flashing at the top of the box (above the "E" of "Event").

3) Right click over the grey area, and select *Show Property View* from the menu that appears (see Figure 12).

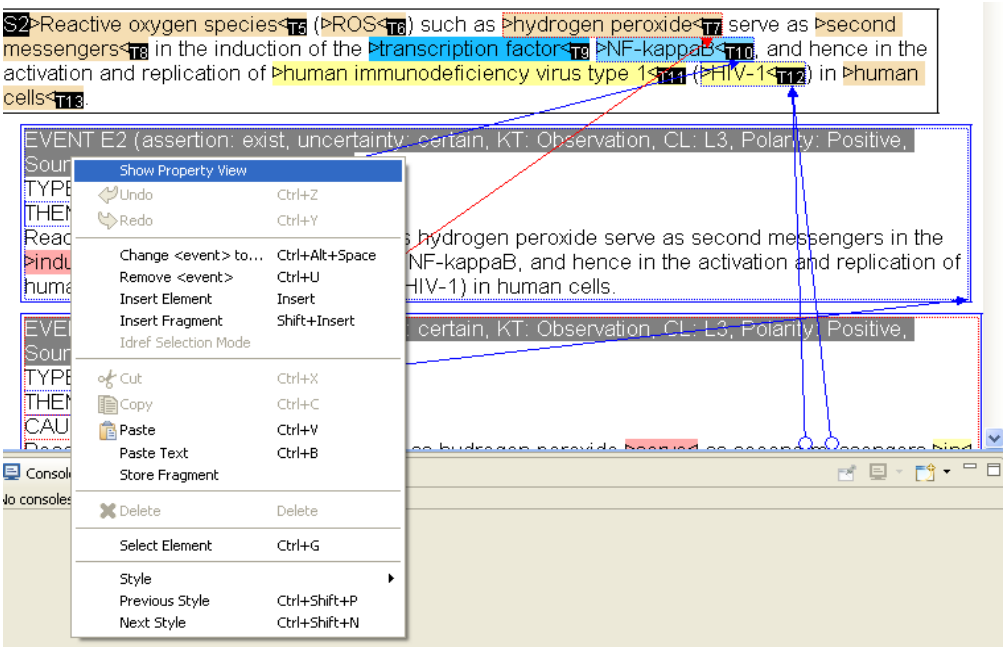

Figure 12 – Preparing to Edit Meta-Knowledge Dimensions

4) A *Properties* window will be displayed, which shows the values of the different annotation dimensions and properties in the form of a table. The names of the dimensions/properties are listed in alphabetical order. The window is shown in Figure 13.

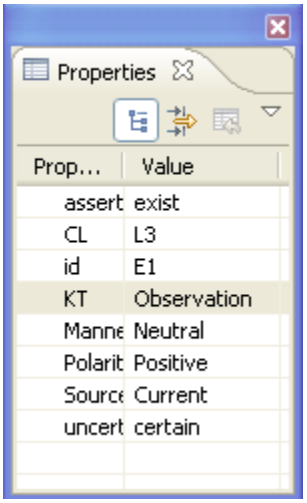

Figure 13 – Properties Window

5) The values of rows corresponding to meta-knowledge dimensions can be edited by clicking over the corresponding row. This will cause an arrow to appear at the right-hand side of the *Value* column. Clicking on this arrow will cause a drop-down menu to appear displaying the possible values for the selected attribute, as shown in figure 14. Selecting a new value from this menu caused the value of the dimension to be changed

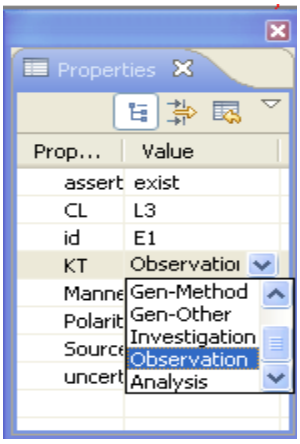

Figure 14 – Selecting an Alternative Dimension Value

6) After the values of all dimensions have been changed as necessary, the "Properties" window can be closed by clicking on the red "X" in the top right hand corner. It should be verified that the values of any dimensions that have been edited have been updated in the grey area within the box.

5.2.4 Annotating Clue Words/Phrases

These are annotated as text span annotations within the *clue* element at the bottom of the event frame annotation box (i.e. the sentence in which the event occurs, and in which other text spans have already been annotated).

Note that if the dimension value is assigned based on features other than specific words/phrases (e.g. the tense of the event trigger word or the position of the sentence within the abstract), then it is *not* necessary to annotate a clue word/phrase. Event trigger words may also act as meta-knowledge clue words (see section 2).

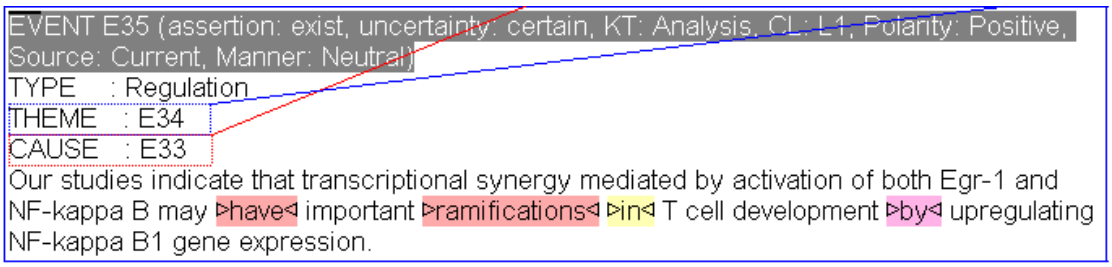

Figure 15 – Sample Event for Clue Annotation

As an example of annotating clues, consider the event shown in Figure 15, for which the KT value of *Analysis* and the CL value of *L1* have been assigned.

Firstly, it is the presence of the word *indicate* that leads to the assignment of the KT value *Analysis*, as it provides the information that the event is based on a conclusion/analysis based on the experimental results.

The annotation of *indicate* as a clue for the assignment of the KT value of *Analysis* proceeds as follows:

1) Drag with the mouse over the word to be annotated
2) Right click with the mouse over the highlighted word
3) Select the item *Insert Element.* This is shown in Figure 16.

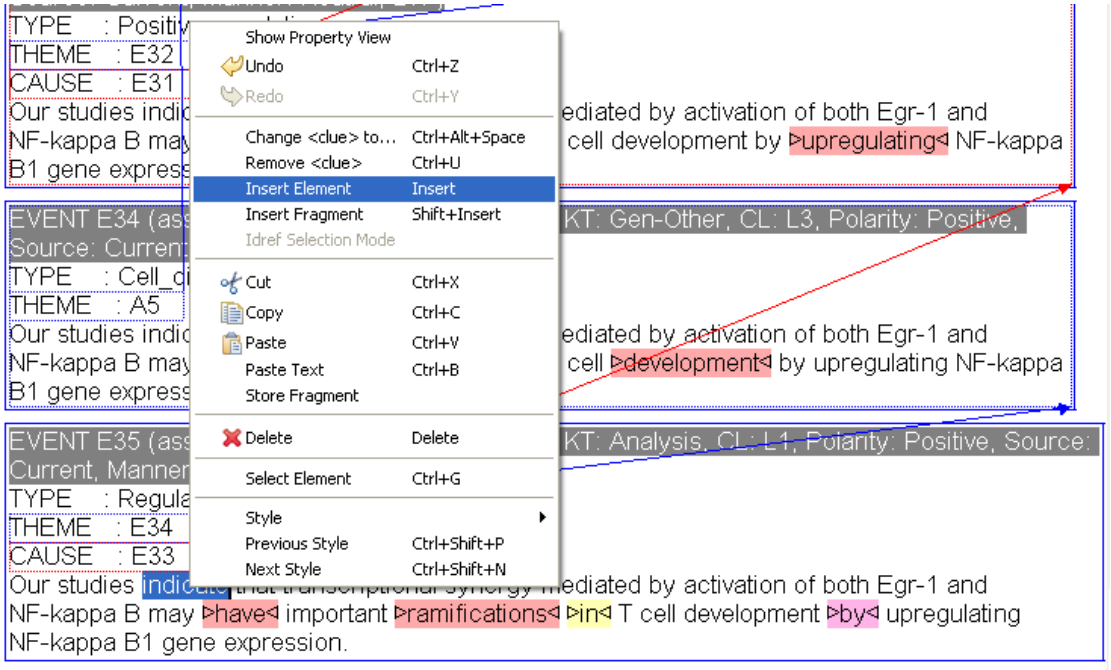

Figure 16 – Inserting a text span annotation

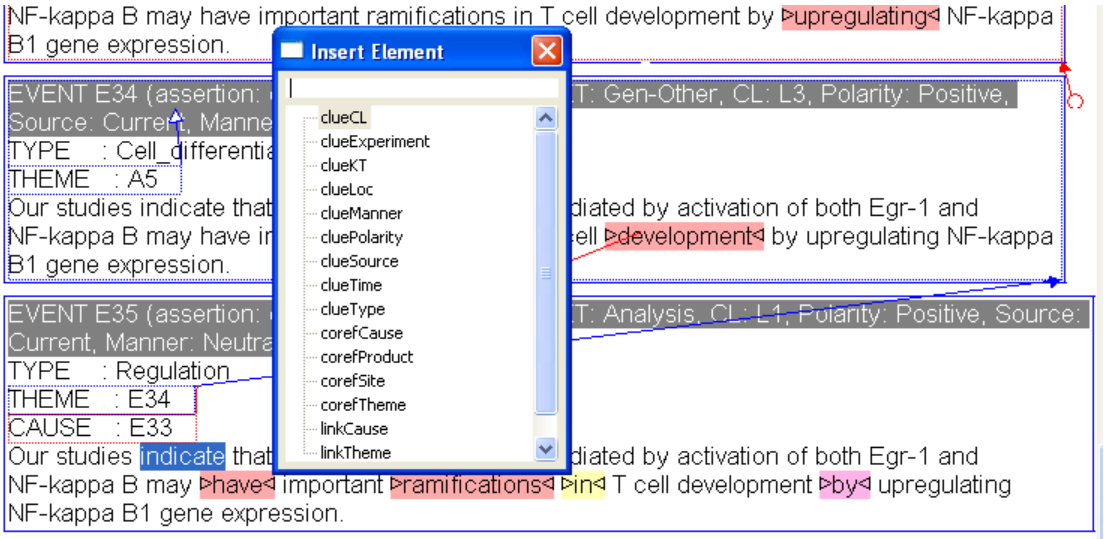

Figure 17 – The "Insert Element" Window

4) An "Insert Element" window will appear, that lists the different categories of text span annotation that can be added. This is shown in Figure 17.

5) The appropriate category should be chosen from this window. In this case of the current example, the correct category to choose is *clueKT*, as this is a clue for the assignment of the Knowledge Type dimension.

6) The newly annotated text span will become highlighted with a background colour according to the category chosen. In the case of *clueKT*, the background colour is indigo. The added annotation is illustrated in Figure 18.

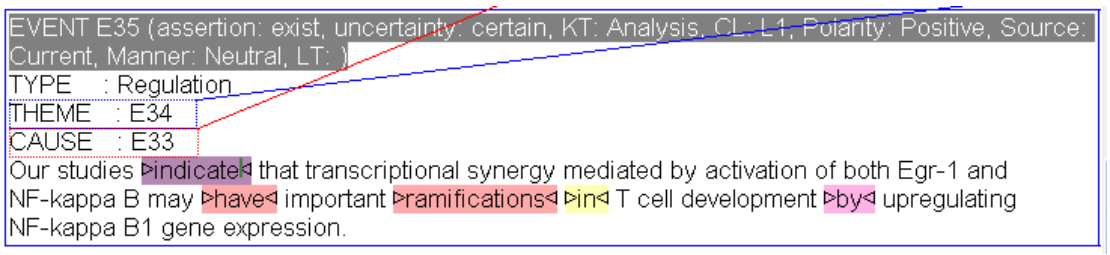

Figure 18 – Event with clueKT marker annotated

The colours of each annotation are as follows:
clueKT – indigo
clueCL – bluish green
cluePolarity –lime green
clueManner – purple
clueSource – red

Returning to the above event, the Certainty Level value of *L1* is assigned on the basis of the presence of the word *may*. This is annotated by following the same steps as above, except that *clueCL* is chosen from the *Insert Element* window. The finished annotated event, complete with meta-knowledge annotation added is shown in Figure 19.

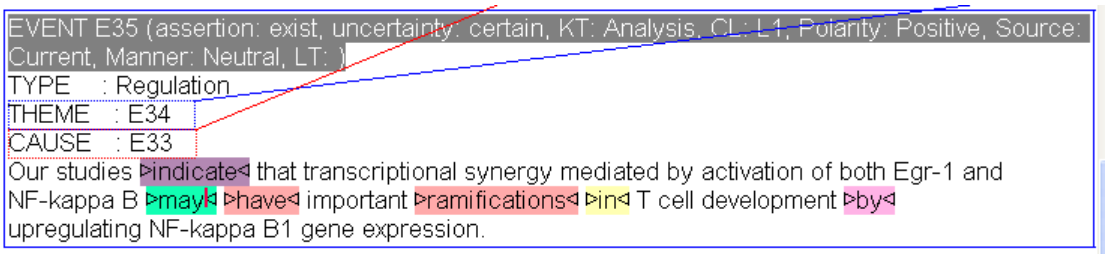

Figure 19 – Event with both clueKT and clueCL markers added

5.3 X-Conc Tips, Pitfalls and Common Sources of Error

5.3.1 Ensuring that the correct annotation is selected

Before performing or editing annotations, it is important to ensure that the correct region on the screen has been selected. This will ensure that no errors occur (e.g., that annotation is carried out for the wrong event).

Before editing values of the meta-knowledge dimensions, you should ensure that you have clicked within the grey area at the top of the appropriate event annotation, and that the cursor line is flashing above the *E* of the word *EVENT*.

5.3.2 Deleting/changing text span annotations

If a meta-knowledge clue word or phrase has been added in error, or if the wrong clue category has been assigned, then the following steps should be followed:

1) Click inside the erroneous annotation. Ensure that the cursor is flashing within the annotation.
2) Right click with the mouse
3) From the menu that appears, choose one of the following options:
    a) *Remove <name_of_annotation>*, e.g. *Remove <clueCL>* to remove the annotation completely
    b) *Change <name_of_annotation> to …*to change the category of the annotation to another one. A window will appear allowing the new category to be chosen.

If the span of the added annotation is incorrect, i.e., if it does not cover the correct number of characters, then the annotation should be removed and added again.

NOTE: Please take care not to delete any text span annotations that were added as part of the original event annotation. If this is done in error, then an *Undo* function is available, either via the right-click menu, the *Edit* menu, or using *CTRL+Z.*

5.3.3 Words and Phrases that are Clues for Multiple Meta-Knowledge Annotations

As explained in section 2, it is possible for certain words and phrases to act as a clue for the assignment of more than one meta-knowledge dimension. The most common occurrences of this phenomenon are words and phrases that jointly denote a Knowledge Type value of *Analysis* as well as a Certainty Level value of L2. Typical markers falling into this category include the verbs *suggest, indicate* and *believe.*

It is possible to create multiple annotations over a single text span, although the annotations must be carried out in the correct order. Only certain combinations of these multiple categories are allowed, according to what we believe to be reasonable combinations.

In order to annotate a word or phase as both a *clueKT* and a *clueCL*, the follwing steps should be taken:

1) The *clueKT* annotation should be added first.
2) The same text span should then be highlighted again, and the right mouse button should be clicked. This time, there will an option on the menu to *Insert <clueCL>*, which should be chosen. As we envisage that only a *clueCL* can occur over the same text span as a *clueKT*, the process of creating this second annotation is somewhat simpler than adding the first.

NOTE: If you wish to create multiple annotations over a single text span, but X-Conc does not allow you to do this, you should contact us to discuss the problem. It may be that there is some combination of clues that we did not consider.

6 Annotation Reference 1: Sequence, Clues and Implications

Annotation Sequence	Dimension	Category	Type of Clue	Implications	
				Current Event	Participant Events
1	Knowledge Type	Investigation	Explicit	CL = L3	KT = Other (unless clearly an *ANALYSIS*)
		Analysis	Explicit	-	KT = Other (Unless the CAUSE is clearly *FACT* or *OBSERVATION*)
		Observation	- Explicit (sometimes) - Implicit (mostly past tense or previous sentence)	CL = L3	KT = Observation or Method EXCEPTIONS: When the semantics of the current event denote that the participant event did not happen. This could be through negation or the meaning of the event trigger. In this case *Other* should be assigned
		Fact	- Explicit (rarely) - Implicit (mostly present tense or previous sentence)	CL = L3	KT = Other (unless clearly a complete fact, in which case Fact may be assigned)
		Method	Explicit (within clueType)	-	-
		Other	Not Annotated	CL = L3	KT = Other
2	Certainty Level	L3	Not Annotated	-	-
		L2	Explicit	KT = Analysis (retrospectively)	KT = Other
		L1	Explicit	KT = Analysis (retrospectively)	KT = Other
3	Polarity	Negative	Explicit	-	KT = Other
		Positive	Not Annotated	-	-
4	Manner	High	Explicit	-	-
		Low	Explicit	-	-

		Neutral	Not Annotated	-	-
5	Source	Other	Explicit	-	-
		Current	Not Annotated	-	-

7 Annotation Reference 2 – List of Typical Clues

<table>
<tr><th>Dimension</th><th>Category</th><th>Typical Clues</th></tr>
<tr><td rowspan="2">Knowledge Type</td><td>Investigation</td><td>
- Verbs in finite form (preceding the event-trigger) or their nominalisations, for example:
<table>
<tr><td>analyze</td><td>compare</td><td>examine</td><td>explore</td></tr>
<tr><td>evaluate</td><td>focus (on)</td><td>investigate</td><td>Study</td></tr>
<tr><td>test</td><td></td><td></td><td></td></tr>
</table>
- Verbs in infinitive form (preceding the event-trigger). This includes all of the above verbs along with some others like:
<table>
<tr><td>ascertain</td><td>define</td><td>elucidate</td><td>identify</td></tr>
<tr><td>determine</td><td>characterize</td><td>distinguish</td><td></td></tr>
</table>
- Please see section 2.1.1 (page 14) for examples
</td></tr>
<tr><td>Analysis</td><td>
- Verbs (finite forms) or their nominalizations preceding the event-trigger, for example:
<table>
<tr><td>appear</td><td>assume</td><td>believe</td><td>conclude</td></tr>
<tr><td>define</td><td>demonstrate</td><td>establish</td><td>evidence</td></tr>
<tr><td>hypothesize</td><td>identify</td><td>indicate</td><td>presume</td></tr>
<tr><td>report</td><td>reveal</td><td>seem</td><td>show</td></tr>
<tr><td>suggest</td><td>contribute</td><td>confirm</td><td>verify</td></tr>
<tr><td>identify</td><td>propose</td><td>corroborate</td><td>realize</td></tr>
<tr><td>postulate</td><td>relate</td><td>detect</td><td>think</td></tr>
</table>
- Conjunctions such as:
<table>
<tr><td>therefore</td><td>thus</td><td>consequently</td><td></td></tr>
</table>
Verbs or nominalizations serving as event-triggers, for example:
<table>
<tr><td>associate</td><td>attribute</td><td>correlate</td><td></td></tr>
<tr><td>implicate</td><td>relate</td><td>CONCLUSION</td><td></td></tr>
</table>
- Modal auxiliaries (if no other Analysis words are present in the sentence):
<table>
<tr><td>could</td><td>may</td><td>might</td><td>can</td></tr>
</table>
- Frequency indicators (if no other Analysis words are present in the sentence):
</td></tr>
</table>

<table>
<tr><td></td><td></td><td>
<table>
<tr><td>frequently</td><td>normally</td><td>occasionally</td><td>often</td></tr>
<tr><td>mostly</td><td>mainly</td><td>usually</td><td></td></tr>
</table>

- Adjectives and adverbs (mostly non-finite verb forms) like:
<table>
<tr><td>capable of</td><td>consistent with</td><td>judged by</td><td>is able to</td></tr>
<tr><td>suggestive of</td><td>potential</td><td>presumably</td><td>apparently</td></tr>
<tr><td>susceptible</td><td></td><td></td><td></td></tr>
</table>
- Please see section 2.1.2 (page 15) for examples</td></tr>
<tr><td></td><td>Observation</td><td>- Explicit word in the same sentence. Typical clue words are:
<table>
<tr><td>detect</td><td>find</td><td>observe</td><td></td></tr>
</table>
- If explicit words are not present, the event trigger verb may provide evidence for the assignment of the Observation category, if it is either:
1) in the past tense
2) in the present tense, but in an observation context
3) A secondary event that is a participant of a primary event assigned the Knowledge Type of Observation
- Please see section 2.1.3 (page 18) for examples</td></tr>
<tr><td></td><td>Fact</td><td>- Events with triggers that describe biological processes in the present tense (could also be Observations according to context). Explicit clue words and phrases are normally not present, with the exception of known, which may sometimes be present.
- Please see section 2.1.5 (page 21) for examples</td></tr>
<tr><td></td><td>Method</td><td>- Any events whose trigger is a word that describes an experimental method. Typical clue words are:
<table>
<tr><td>addition</td><td>incubated</td><td>pretreated</td><td>stimulation</td></tr>
</table>
- Please see section 2.1.4 (page 20) for examples</td></tr>
<tr><td></td><td>Other</td><td>- Secondary events whose primary event has the Knowledge Type of Analysis, Investigation or Fact.
- Secondary events whose primary event has been negated (i.e., Polarity = Negative).
- Secondary events whose primary event has the Knowledge Type of Observation, where the meaning of the trigger verb of the primary event conveys the fact that the secondary event did not take place. Examples of such clue words include inhibit and suppress etc.
<u>NOTE:</u> Other secondary events whose primary event has the Knowledge Type of Observation would also normally</td></tr>
</table>

<table>
<tr><td></td><td></td><td>be assigned the Knowledge Type of Observation
- Events that describe properties of entities
- Default category i.e., if no other category is applicable.
- Please see section 2.1.6 (page 22) for examples</td></tr>
<tr><td rowspan="3">Certainty Level</td><td>L3</td><td>- Default category i.e., if no other category is applicable.</td></tr>
<tr><td>L2</td><td>- Probability indicators are:
<table><tr><td>likely</td><td>probably</td><td>can</td><td>presumably</td></tr><tr><td>able</td><td>ability</td><td>susceptible</td><td>evidence</td></tr></table>- Analysis verbs such as:
<table><tr><td>believe</td><td>hypothesize</td><td>indicate</td><td>suggest</td></tr><tr><td>assume</td><td>seem</td><td>appear</td><td>suspect</td></tr><tr><td>propose</td><td>implicate</td><td>postulate</td><td>think</td></tr></table>- Frequency indicators like:
<table><tr><td>normally</td><td>frequently</td><td>mostly</td><td>mainly</td></tr><tr><td>usually</td><td></td><td></td><td></td></tr></table>- Please see section 2.2.2 (page 24) for examples</td></tr>
<tr><td>L1</td><td>- Modal auxiliaries and possibility indicators like, , and etc.
<table><tr><td>possibly</td><td>may</td><td>might</td><td>perhaps</td></tr><tr><td>unclear</td><td>potentially</td><td></td><td></td></tr></table>- Analysis verbs such as:
<table><tr><td>speculate</td><td></td><td></td><td></td></tr></table>- Frequency indicators like:
<table><tr><td>rarely</td><td>scarcely</td><td>sometimes</td><td></td></tr></table>- Please see section 2.2.3 (page 26) for examples</td></tr>
<tr><td>Polarity</td><td>Negative</td><td>- NOTE: This is a fairly large list of words which could <u>potentially</u> denote negative polarity, given the correct context. If you encounter one of these words, please take extra care to ensure that negative polarity is indeed being described.
- The adverbial not and the nominal no.
<table><tr><td>no</td><td>not</td><td>nor</td><td></td></tr></table>- Verbs like:
<table><tr><td>fail</td><td>lack</td><td>loss</td><td>impair</td></tr></table></td></tr>
</table>

<table>
<tr><td rowspan="3"></td><td rowspan="2"></td><td>
<table>
<tr><td>prevent</td><td></td><td></td><td></td></tr>
</table>
- Adjectives like:
<table>
<tr><td>independent</td><td>absent</td><td>barely</td><td>cannot</td></tr>
<tr><td>deficient</td><td>unable</td><td>inactive</td><td>insensitive</td></tr>
<tr><td>insufficient</td><td>limited</td><td>negative</td><td>resistant</td></tr>
<tr><td>unaffected</td><td>unchanged</td><td>defective</td><td></td></tr>
</table>
- Adverbs like:
<table>
<tr><td>without</td><td>independently</td><td>instead</td><td>neither</td></tr>
<tr><td>never</td><td></td><td></td><td></td></tr>
</table>
- Nouns like:
<table>
<tr><td>exception</td><td>absence</td><td>deficiency</td><td>failure</td></tr>
<tr><td>inability</td><td>resistance</td><td>none</td><td></td></tr>
</table>
- Prepositions like:
<table>
<tr><td>except</td><td>without</td><td></td><td></td></tr>
<tr><td></td><td></td><td></td><td></td></tr>
</table>
Please see section 2.3.2 (page 26) for examples</td></tr>
<tr><td>Positive</td><td>- Default category i.e., if no other category is applicable.</td></tr>
<tr><td rowspan="3">Manner</td><td>High</td><td>- Adverbs and adjectives like:
<table>
<tr><td>markedly</td><td>rapid</td><td>rapidly</td><td>severe</td></tr>
<tr><td>significant</td><td>significantly</td><td>strong</td><td>strongly</td></tr>
<tr><td>potent</td><td>high</td><td>considerable</td><td></td></tr>
</table>
- Please see section 2.4.1 (page 28) for examples</td></tr>
<tr><td>Low</td><td>- Adverbs and adjectives like:
<table>
<tr><td>barely</td><td>limited</td><td>little</td><td>low</td></tr>
<tr><td>lower</td><td>weak</td><td>modest</td><td></td></tr>
</table>
- Please see section 2.4.2 (page 29) for examples</td></tr>
<tr><td>Neutral</td><td>- Default category i.e., if no other category is applicable.</td></tr>
<tr><td>Source</td><td>Other</td><td>- Phrase such as previous studies and previously etc.
<table>
<tr><td>previous study/studies/report(s)</td><td>previously</td></tr>
</table></td></tr>
</table>

		recent study/studies/report(s) ***recently*** - Citations - Please see section 2.5.2 (page 32) for examples
	Current	- Default category i.e., if no other category is applicable.