Categorising Modality in Biomedical Texts

Paul Thompson[1], Giulia Venturi[2], John McNaught[1], Simonetta Montemagni[2] and Sophia Ananiadou[1],

[1]National Centre for Text Mining, University of Manchester, UK

[2]Istituto di Linguistica Computazionale, CNR (Italy)

E-mail: {paul.thompson, john.mcnaught,sophia.ananiadou}@manchester.ac.uk,
{giulia.venturi,simonetta.montemagni}@ilc.cnr.it

Abstract

The accurate recognition of modal information is vital for the correct interpretation of statements. In this paper, we report on the collection a list of words and phrases that express modal information in biomedical texts, and propose a categorisation scheme according to the type of information conveyed. We have performed a small pilot study through the annotation of 202 MEDLINE abstracts according to our proposed scheme. Our initial results suggest that modality in biomedical statements can be predicted fairly reliably though the presence of particular lexical items, together with a small amount of contextual information.

1. Introduction

Text processing systems tend to focus on factual language (Hahn & Wermter, 2006; McNaught & Black, 2006). However, modality is a common phenomenon which must be taken into account to correctly interpret text. Modality is concerned with the opinion and attitude of the speaker (Lyons, 1977). Palmer (1979) distinguishes three types of modality: *epistemic* (making judgements about the truth of a proposition), *deontic* (concerned with permission) and *dynamic* (concerned with the potential of a situation to occur).

Our concern here is *epistemic* modality in biomedical text, which covers the expression of the author's level of confidence towards a proposition, but may also indicate the type of knowledge, assumptions or evidence on which the proposition is based (Coates, 1995).

It also covers *speculation.* Light et al. (2004) and Medlock & Briscoe (2007) show that successful classification of biomedical sentences for speculation depends on the presence or absence of speculative cue words or phrases. Whilst more complex syntactic contexts, e.g. conditional clauses, are a possible way express modality in texts (Sauri et al, 2006), corpus-based studies of *hedging* (i.e. speculative statements) in biological texts by Hyland (1996a, 1996b) reinforce the above experimental findings: 85% of hedges were realised lexically, rather than through more complex means.

Previous efforts at annotating modal information in biomedical texts (e.g. Light et al., 2004; Wilbur et al. 2006; Medlock & Briscoe, 2007) have been at the sentence or sentence fragment level only, without explicit annotation of the modal cue words. Given the importance of these cue words, a list of modal lexical items used within the field, categorised according to the information they express, would be a useful resource for the automatic processing of biomedical texts.

Previous lists (e.g. Hyland, 1996a; Rizomiliti, 2006) suffer from being either incomplete or coarse-grained. Here, we describe the collection and multi-dimensional classification of a preliminary set of words and phrases that express modality within biomedical texts. We then report on initial validation of our classification, via annotation of modal information that modifies previously-annotated gene regulation events in a small corpus of MEDLINE abstracts.

Although oriented towards biological events, our proposed categorisation could be equally valid for other applications, e.g. helping to determine intent of citations, as suggested by DiMarco & Mercer (2004).

2. Modality in Scientific Texts

Hyland (1996a; 1996b) shows that modals such as *may, could, would* etc., play a relatively minor role in expressing modality in biological texts. This is evident when the proportions of word categories occurring in such texts are compared to those calculated by Holmes (1988) on general academic articles from the Brown corpus of American English and the LOB corpus of British English. See Table 1.

	Hyland (Biology)	Holmes (gen. academic)
Lexical verbs	27.4%	35.9 %
Adverbials	24.7%	12.8 %
Adjectives	22.1 %	6.6 %
Modal verbs	19.4 %	36.8 %
Nouns	6.4 %	7.7 %

Table 1: Comparison of modal items in different text types

It is the lexical (i.e. non-modal) verbs, adjectives and adverbs that dominate in expressing modality in biological research articles. Thus, we collected a set of modal words and phrases that are relevant within the biomedical domain.

3. Collecting Modal Items from Biomedical Texts

Rizomilioti (2006) provides a comprehensive base list of modal lexical items drawn from academic research

articles in biology, archeology and literary criticism (200,000 words each).

As we hypothesised that modal lexical items can vary amongst text genres, we eliminated any items with no modal sense within the biomedical literature.

113 abstracts were taken from a corpus of MEDLINE abstracts on E. Coli (approximately 30,000 abstracts). Within these, any additional lexical markers of modality not present on Rizomilioti's list were identified. Examples included *evidence, observe, predict, imply, be consistent with* and *potential.*

For each item in the resulting combined list, we calculated its frequency within the complete E. Coli corpus, and discarded 26 items occurring fewer than 10 times (9 had zero occurrence). Discarded words included those indicating a high degree of confidence (e.g. *surely, patently, admittedly*, *attest, emphasise*) and those expressing doubt (e.g. *alledgedly, improbable, doubtful, ostensibly*).

We examined the usage of each remaining word in context within the corpus using a concordancer from the Multilingual Corpus Toolkit[1], and discarded any words not having a modal meaning in any contexts within the corpus. Examples included the verb *stress* and the adverb *naturally,* both of which Rizomilioti judged to indicate a high degree of confidence. In our corpus, *stress* almost always occurs as a noun, e.g. *oxidative/environmental stress*, whilst *naturally* most commonly occurs in phrases such as *naturally occurring.* Following this step, 90 lexical items remained. Table 2 shows the most frequently occurring of these in the E. Coli corpus, which correspond well with the highest ranked terms identified by Hyland and Rizomilioti.

show (17836)	may (5826)
suggest (11850)	demonstrate (4817)
indicate (8511)	reveal (4467)
observe (6177)	could (4247)
identify (5494)	appear (4212)

Table 2: Most frequent modal words in the E.Coli corpus

4. Classifying Modality in Scientific Texts

To propose a categorisation model for our list of modal lexical items, we considered a number of existing models and annotation schemes. Light et al. (2004) and Medlock & Briscoe (2007) are concerned primarily with the speculative/non-speculative distinction; other models are more complex and include multiple "dimensions". Rubin et al. (2005) annotate certainty in newspaper articles along 4 separate dimensions: a) degree of *certainty*; b) *focus,* i.e. whether the statement is abstract (opinions, beliefs, assessments) or factual; c) *perspective,* i.e. the writer's or a reported point of view; and d) *time,* i.e. whether the reported event is in the past, present or future. Lexical markers are explicitly annotated to give evidence for the value assigned to each attribute, suggesting that separate sets of words or phrases are used to express these different dimensions.

Although newspaper articles are very different from biomedical texts, the *certainty, focus* and *perspective* dimensions also seem relevant for us. The following corpus sentence illustrates the possibility of identifying these different dimensions through separate words or phrases:

We suggest that these two proteins may form a complex in the membrane which acts at late steps in the export process

The word *we* shows that the perspective is the authors' own, *suggest* provides *focus* information (i.e. this is a speculation rather than a definite fact) whilst *may* conveys the author's level of certainty.

4.1 Evidence Underlying Statements

An aspect not covered by Rubin et al.'s model, and yet highly relevant in scientific literature, is the source or type of *evidence* underlying a statement. The importance of this within the biomedical field is illustrated in annotation using the Gene Ontology (GO) (Ashburner et al, 2000). This requires gene associations to be attributed to the literature through the assignment of *evidence codes*, which denote the type of evidence available, e.g. experimental evidence, evidence through citations, or evidence inferred by the curator from other GO annotations.

Wilbur et al.'s annotation scheme also uses *evidence*; some of the possible values correspond closely to the main evidence categories used by GO curators, thus reinforcing that this type of information is important to domain experts. Sentences or sentence fragments are annotated for evidence as follows: a) no evidence, b) claim of evidence without verifying information, c) citation of other papers or d) explicit reference to experimental findings or results described within the paper.

4.2 Interpretation of Evidence

Certain verbs, like *see, indicate* and *find,* can help to identify statements containing reference to evidence. Wilbur et al.'s scheme determines the value of the *evidence* attribute largely from the type of subject taken by the verb, or the presence of citations. A subject such as *our results* provides explicit reference to results within the paper, whilst *previous studies* makes a claim of evidence, which may or may not be backed up by a citation.

Whilst the *type* of evidence behind a statement is important within the domain, another relevant type of information is how that evidence is to be *interpreted.* The choice of verb (or other modal lexical item) is important for this: whilst a statement beginning "*we **see** that …*" normally expresses an observation based on experimental findings or results, a sentence of the form "*Previous experiments **indicate** that…*" would imply that reasoning has taken place to arrive at the statement that follows.

Palmer's (1986) model for the 4-way classification of non-factual statements takes such distinctions into account, yielding *speculative, deductive* (derived from inferential reasoning or conclusions), *quotative* (specifying and acknowledging previous findings) and *sensory* (referring to apprehending, sensing or observing).

[1] http://personalpages.manchester.ac.uk/staff/scott.piao/research/

This model has been analysed for biological texts by Hyland (1996a). It has similarities with Wilbur et al.'s *evidence* attribute, in that the *quotative* category encompasses statements that cite other works. However, other types of statements backed by evidence are divided into *sensory* and *deductive,* according to whether they are based on observations or reasoning. Hyland's examples suggest that lexical items themselves can be used to distinguish between the *speculative, deductive* and *sensory* categories. For example, *appear* and *seem* are sensory verbs, whilst the verbs *propose, believe* and *speculate* fit well into the *speculative* category. Likewise, the verbs *infer, indicate* and *imply* are typical indicators of the *deductive* category.

5. Proposed Categorization Scheme

We conclude that the following factors are important to the interpretation of statements in scientific literature:

a) whether the statement is a speculation or based on factual data (e.g. experimental findings or results)
b) the type/source of the evidence
c) the interpretation of the evidence
d) the level of certainty towards the statement

We take Palmer's model as a starting point for our own proposed categorisation, as it covers the above factors a), b) and c), at least to a certain extent. Ad factor a), *sensory, deductive* and *quotative* statements are normally based on factual data, whilst speculations fall into the *speculative* category. Ad factor b), Palmer's categories allow different types of evidence to be distinguished. So, a *speculative* statement is not normally backed by evidence, whilst *sensory* and *deductive* statements would normally contain claims of evidence or reference to experimental findings. Meanwhile, *quotative* statements normally provide evidence through citation of other papers. Finally, ad factor c), different *interpretations* of evidence may be distinguished according to whether the statement is *sensory* or *deductive*.

Arguably, Palmer's categories implicitly encode certainty level information. A speculation is, for example, normally a less confident assertion than one backed by evidence of some sort. However, this does not necessarily follow: deductions and experimental observations may be made with varying degrees of confidence through the use of explicit certainty markers like *may* or *probably*. Thus, we follow Rubin et al. (2005) and Wilbur et al. (2005), in categorising certainty level as a separate dimension.

We further observed that *quotative* does not form a distinct category of statements. Consider: "*Trifonov [38] has suggested that*…". Here, the cited work speculates about the statement that follows, and so the sentence is *both* quotative *and* speculative. We thus classify the *point of view* of the statement (i.e. that of the author or a cited work) as a separate dimension. Whilst this does not correspond to modal information per se, its identification is important for correct interpretation of certain sentences containing modal lexical items, e.g. in determining the source of evidence presented. A sentence beginning O*ur data implies that* …is the author's point of view, indicating that the experimental findings discussed are drawn from the current paper rather than another source.

Our categorisation scheme for modality in biomedical texts thus consists of the following 3 "dimensions" of information:

1) *Knowledge Type,* encoding the type of "knowledge" that underlies a statement, encapsulating both whether the statement is a speculation or based on evidence *and* how the evidence is to be interpreted.
2) *Level of certainty,* indicating how certain the author (or cited author) is about the statement.
3) *Point of View*, indicating whether the statement is based on the author's own or a cited point of view or experimental findings.

Recognition of the *Point of View* level is aided through finding strings such as *we* and *our* (corresponding to the author's point of view), or various forms of citations for cited points of view. According to our scheme, the possible values for the *Point of View* dimension are *writer* or *other.* The other two dimensions can be recognised largely through the presence of lexical items such as the ones collected from our corpus of E. Coli abstracts.

5.1 Knowledge Type

The majority of lexical items within our list have been categorised under *Knowledge Type*. Three of the subclasses we use are taken from Palmer's model: *speculative, deductive* and *sensory.* To these, we add a fourth category of words whose members explicitly mark a statement as *describing* experimental results or findings, rather than observations or deductions made from them. Such statements are marked by words such as *show, reveal, demonstrate* or *confirmation.* As experiments are normally carried out to prove or demonstrate a hypothesis, we label this class of words *demonstrative.*

The largest category of items is the *speculative* one, containing 30 members from our preliminary list. These include not only verbs or their nominalised equivalents such as *predict, prediction, hypothesize, hypothesis,* etc., but also other nouns such as *view* and *notion,* adjectives like *conceivable* and phrases such as *in theory* and *to our knowledge.* Other categories are smaller, ranging from 8-12 items, consisting mainly of verbs and nominalised forms. So, a *deductive* statement can be denoted by *interpret, indication* or *deduce,* whilst sentences with *sensory* evidence can be marked with words such as *observation, see* or *appear.*

However, context may be required to correctly determine the category of statements denoted by certain *Knowledge Type* words: *suggest,* when used with a human subject, e.g. *We suggest …* or in the passive voice, e.g. *It is suggested…*, denotes a speculation; however, when the subject is inanimate, e.g. *The results suggest …,* there is an implication that a deduction has been carried out.

5.2 Level of Certainty

The partitioning of lexical items or statements into various degrees of certainty has been extensively studied, but little consensus has been reached. Rubin (2007) notes an ongoing discussion about whether they should be arranged on a continuum or into discrete categories.

Hoye (1997) proposes that there are at least three articulated points on the epistemic scale: *certainty,*

probability and *possibility*. However, recent works have suggested more fine-grained partitions, with either 4 distinct levels (Rubin et al, 2005; Wilbur et al. 2006) or even 5 levels (Rubin, 2007). Annotation experiments according to this 5 level system (i.e. *absolute certainty, high certainty, moderate certainty, low certainty* and *uncertainty*) suggested that English may not be precise enough to distinguish so many shades between certainty and doubt. However, a 4-level distinction appears more feasible, with successful application in both the newspaper (Rubin et al., 2005) and biomedical domains (Wilbur et al., 2006). In the latter case, inter-annotator agreement rates of approximately 80% were reported.

Thus, we derived a four-way classification of lexical items denoting certainty: *Absolute*, *High*, *Moderate* and *Low.*

Within the scientific literature, a statement marked as *known* is normally an accepted fact within the field, and so is assigned the *Absolute* certainty value. Statements marked with words such as *probable, likely* or *clearly* express a *high* degree of confidence. Words such as *normally* and *generally* are also placed in this category, denoting that a specified event takes place most of the time, and thus expressing a high degree of confidence that the statement is true. Also within the *High* category are words and phrases that only express certainty when combined with certain *Knowledge Type* items. *Strongly* can be used in sentences of the following form: *The results strongly suggest that ….* Here, *suggest* has a *deductive* meaning, and *strongly* indicates a high degree of confidence towards this deduction. Words and phrases such as *support, in agreement with* and *consistent with* can be used with speculative nouns (e.g. *theory, notion* or *view*) to lower the speculation (and hence increase the certainty) of the statement.

Moderate items specify a more "neutral" certainty level, without strong indication of whether the statement is more likely to be true than false. Examples include *possibly* and *perhaps,* as well as some modal auxiliary verbs like *may* and *could.* Finally, *low* certainty level items have more negative undertone, signaling little confidence in the statement they modify, e.g. *unlikely*.

6. Testing the Classification Scheme

Our work has been carried out in the context of the BOOTStrep project (FP6 - 028099), aimed at building a bio-lexicon and bio-ontology for biomedical text mining. As part of the project, we have been creating a corpus of E. Coli abstracts annotated with *gene regulation* bio-events (Thompson et al., 2008). Events are centred around verbs (e.g. *regulate*) or nominalised verbs (e.g. *expression)*, and event annotation consists of identifying and classifying the semantic *arguments* or *participants* in the event. Note that event annotation was carried out on top of shallow parsed (pos-tagged and chunked) texts[2]: the advantages of such a choice range from practical ones, i.e. annotated corpora can be produced with much less work, to more substantial ones, i.e. previous levels of annotation can drive the annotation process, thus resulting in an increase in efficiency and consistency for any new annotation.

From the annotated events, patterns (i.e. *semantic frames*) relating to the behaviour of each verb and nominalised verb can be learnt and included within the bio-lexicon; these can help in the automatic extraction of facts from biomedical texts. As the annotated events correspond to facts of biomedical interest, we considered them a useful starting point for the verification of our proposed modality classification.

Thus, we carried out a small experiment, in which modality was annotated within a small set (i.e. 202) of these event-annotated abstracts, using *WordFreak,* a Java-based linguistic annotation tool (Morton & LaCivita, 2003), which was customized to the task.

Due to the linguistically-driven purposes as well as the small size of the corpus exploited in this feasibility study, annotation was carried out by a single annotator with linguistic expertise. However, extensive support was provided by two researchers, one with a background in linguistics, and the other one in biology, to discuss open issues raised during the annotation process in order to improve the semantic stability and reliability of the annotations produced.

6.1 Annotation process

Each sentence containing a previously-annotated gene regulation event was studied, and modality annotation was performed only on those sentences in which the description of the event contained explicit expression of modal information: modal information was only annotated if it was within the scope of the gene regulation event described. Let us consider, for example, the *derepress* bio-event, described in the sentence "*We suggest that overproduction of SlyA in hns(+) E. coli derepresses clyA transcription by counteracting H-NS"*, which was annotated as follows:

VERB: *derepresses*
AGENT: *overproduction*
THEME: *clyA transcription*
MANNER: *counteracting*

The modality annotation process started from the event anchor, i.e. the verb *derepress*. Words or phrases expressing modal information and linguistically bound to the event anchor were searched for within the sentence's span. If such items were found, values from the proposed sets were selected for one or more of the three dimensions of the annotation scheme, i.e. *Point of View, Knowledge Type* and *Certainty Level*. For the *Knowledge Type* and *Certainty Level* attributes, a value was only selected if there was *explicit* lexical evidence in the sentence. In the case at hand, *suggest* was annotated as the lexical modality marker conveying information about *Knowledge Type*, whose associated value is *deductive*. The word *We* was interpreted as lexical evidence that the reported *Point Of View* was that of the writer.

Each piece of lexical evidence (i.e. lexical modality marker) could only be used to assign a value to *one* of the annotation dimensions. Thus, it was not possible to use a single word or phrase to assign values to both the *Knowledge Type* and *Certainty Level* dimensions.

If one or both the *Knowledge Type* or *Certainty Level* attributes were assigned, the *Point of View* attribute was

[2] Each abstract to be annotated with gene regulation bio-events was first pre-processed with the GENIA tagger (Tsuruoka et al, 2005).

also instantiated. If no explicit lexical evidence was available for the assignment of this attribute, a "default" value of *writer* was assigned, i.e. it was assumed that the *Point of View* was expressed implicitly.
The annotator used the preliminary categorisation of modal lexical items as a starting point for the annotation of the *Knowledge Type* and *Certainty Level* attributes, although she was not bound by this categorisation, nor was her annotation limited to only those items on the list: part of the purpose of the annotation was to discover the semantic stability of the lexical items within our proposed categories, as well as to discover other modality markers missing from the preliminary list.

7. Results

The 202 MEDLINE abstracts annotated for modal information contained a total of 1469 gene regulation events. 249 of these events (i.e. 16.95%) were annotated with modality information. Table 3 shows general statistics about the *dimensions* of the modal markers that were present in the description these events, whilst Table 4 shows the distribution of the annotations amongst the various values within each dimension of the scheme.
The number of modality annotations may at first seem rather low, with an average of 1.31 annotations per abstract. However, a number of points should be noted. Firstly, lexical markers of modality are generally quite sparse within texts. Secondly, as pointed out above, modality annotations have only been carried out on top of previously annotated bio-events, and there was an average of 6.05 bio-event annotations per abstract. Rather than aiming to annotate *all* modal information expressed within the abstracts, our case study is firstly aimed at verifying whether the modality classification scheme is suitable for a corpus of biomedical texts, and secondly, it is focused on the discovery of the main domain-relevant problems and features involved, as well as clues which can drive future work.
There follows a number of annotation examples. In each case, the modality marker(s) and the Point Of View marker (if present) have been underlined, with the corresponding category placed in brackets. The verb which forms the focus of the associated bio-event is emboldened.

a) *Therefore, we* [WRITER] *suggest* [DEDUCTIVE] *that overproduction of SlyA in hns(+) E. coli* ***derepresses*** *clyA transcription by counteracting H-NS.*
b)*We* [WRITER] *have shown* [DEMONSTRATIVE] *that the open reading frame ybbI in the genomic sequence of Escherichia coli K-12* ***encodes*** *the regulator of expression of the copper-exporting ATPase, CopA*
c)*We* [WRITER] *speculate* [SPECULATIVE] *that the product of this gene is* ***involved*** *in the attachment of phosphate or phosphorylethanolamine to the core and that it is the lack of one of these substituents which results in the deep rough phenotype.*

A single modality marker may also express the same information relative to more than one bio-event in the case of a coordinated structure, e.g. :

Band shift experiments showed [DEMONSTRATIVE] *that AllR* ***binds*** *to DNA containing the allS-allA intergenic region and the gcl(P) promoter and its binding is* ***abolished*** *by glyoxylate.*

Modal marker(s) present	Count	% of total events
Knowledge Type only	192	77.11 %
Certainty Level only	40	16.07%
Knowledge Type + Certainty Level	17	6.83%

Table 3: Distribution of modality markers within annotated events

Dimension	Value	Count	% of annotations within dimension
Knowledge Type	DEMONSTRATIVE	110	52.63%
	DEDUCTIVE	56	26.79%
	SENSORY	25	11.96%
	SPECULATIVE	18	8.61%
Certainty Level	ABSOLUTE	4	7.01%
	HIGH	15	26.31%
	MODERATE	34	59.64%
	LOW	2	3.50%
Point Of View	WRITER	213	92.20%
	OTHER	18	7.79%

Table 4: Distribution of modality annotations within the different dimensions

7.1 *Knowledge Type* Information

The *Knowledge Type* dimension is the most frequently annotated (77.11% of annotations). The most common value for this dimension is *demonstrative* (52.63% of Knowledge Type annotations), whilst the least widespread type of knowledge is *speculative* (8.61%).
These statistics are perhaps unsurprising, given that the current pilot study has been carried out on abstracts. *Demonstrative* events are explicitly marked as describing experimental results, particularly those which prove hypotheses or predictions. These are exactly the sorts of events that we can expect to occur most frequently in abstracts; within the short amount of space available, authors normally aim to emphasize the definite results that their experiments have produced.
The annotation experiment has highlighted a potential need to add an additional value for the *Knowledge Type* dimension. Consider the following examples:

a) *The model states that the lex (or exrA in E. coli B) gene* ***codes*** *for a repressor.*
b) *Mutations in yjfQ allowed us to identify this gene as the regulator of the operon yjfS-X (ula operon), reported to be* ***involved*** *in L-ascorbate metabolism.*

Events that are introduced by verbs such as *state* or *report* do not fit well into one of our other four *Knowledge Type* categories. They are used to introduce facts, either cited from previous work or earlier in the paper, but without taking a particular stance to them, i.e. there is no speculation or deduction involved, and there is no reference to active proof or demonstration that an

assertion or hypothesis is true.
Statements such as the above fit into Hyland's (1996a) description of the *quotative* category, i.e. specifying and acknowledging previous findings. Thus, the *quotative* label can apply to a wider range of statements than just those that contain citations. Therefore, we propose to introduce the *quotative* category into our classification as a further *Knowledge Type* category to cover statements that specify or acknowledge previous findings through explicit lexical items.
Our annotation also revealed that, whilst the majority of *Knowledge Type* items are fairly stable semantically within their assigned categories, a small number of items do not fit neatly within a single category. The verb *seem* was originally placed within the *sensory* category, following Hyland. However, there is often a speculative aspect to its meaning, as confirmed by Dixon (2005): *seem* is used "when there is not quite enough evidence" (p. 205). The degree of speculation conveyed may vary according to the context: this is an area for further research.

7.2 *Certainty* Information

Certainty level markers are considerably less common than *Knowledge Type* markers, representing 16.07% of the modality annotations. The most widespread value among these annotations is *moderate* (59.64%).
The high percentage of *moderate* markers can again be explained by the text type, i.e. abstracts. The results concerning *Knowledge Type* illustrated that *demonstrative* statements are most common: authors are keen to emphasize the experimental results that they have produced. If there is doubt about these results, this can be indicated thought an explicit certainty level marker. A *moderate* (and hence neutral) certainty level marker may be the "safest" choice here.
Certainty Level markers occur most commonly without an accompanying *Knoweldge Type* marker, as in:

*EvgA is likely [HIGH] to directly **upregulate** operons in the first class, and indirectly upregulate operons in the second class via YdeO.*

As mentioned previously, *Knowledge Type* markers implicitly encode certainty level information. Thus, when a statement is explicitly marked as a speculation or deduction, the use of an explicit marker of certainty may be unnecessary, except for emphasis, or to alter the "default" certainty level associated with the *Knowledge Type* item.
Nevertheless, our annotation has served to identify a small number of cases (6.83%) that contain explicit markers of both *Knowledge Type* and *Certainty Level* information. Such cases provide evidence that our proposed separate dimensions of annotation are indeed well motivated. Some examples are shown below:

a) *No reverse transcriptase PCR product could be detected for hyfJ-hyfR, suggesting* [DEDUCTIVE] *that hyfR-focB may* [MODERATE] *be independently **transcribed** from the rest of the hyf operon.*
b) *We* [WRITER] *suggest* [SPECULATIVE] *that these two proteins may* [MODERATE] ***form** a complex in the membrane which acts at late steps in the export process.*

A large number of certainty level markers are fairly stable in terms of semantics, particularly adjectives and adverbs such as *probable, possibly* or *likely.* Another category of words that play a central role in expressing certainty in our corpus is the modal auxiliaries (e.g. *can, may* or *could*), which represent 40.35% of the total number of *Certainty Level* markers. However, their interpretation is more problematic than adjectives and adverbs like those listed above. In general, *can, may* and *could* can have the following senses:

1) *Moderate* level of certainty
2) Theoretical possibility (indicating that an event has the potential to occur)
3) Ability
4) Permission

Whilst the *permission* sense is rarely relevant within biomedical texts, examples of the other three senses can be readily identified within our corpus. Some examples involving *may* are shown below:

1) **Certainty level marker**
The DNA-binding properties of mutations at positions 849 and 668 may [MODERATE] *indicate* [DEDUCTIVE] *that the catalytic role of these side chains is **associated** with their interaction with the DNA substrate.*
2) **Theoretical possibility marker**
*The expression of nifC may be **coregulated** with nitrogen fixation because of the presence of nif-distinctive promoter and upstream sequences preceding nifC-nifV omega-nifV alpha.*
3) **Ability marker**
*Results obtained indicate that the nrdB gene has a promoter from which it may be **transcribed** independently of the nrdA gene.*

Thus, the presence of these modal auxiliaries does not guarantee that certainty level is being conveyed. Determining the correct sense can be a difficult task, which requires in-depth knowledge of the domain, and often requires examining a wider context than just the sentence itself.
Whilst this could prove problematic in the automatic recognition of modality, Collins (2006) suggests that for each verb, one sense is usually more likely than the others. In his study of *can* and *may* in various spoken and written sources, he found that *may* was used as a certainty level marker in 83.5% of cases, whilst only 1.1% of occurrences of *can* concerned certainty level. A default interpretation of each modal could thus be used. Further study of the context of these items may reveal clues that could determine when a non-default value should be assigned.
Our studies have shown that the meaning of *can* mainly corresponds to the "ability" sense, although "theoretical possibility" is also possible, as shown in the following examples:

a) *The enhanced expression of tac-dnaQ reduces 10-fold the frequency of UV-induced Su+ (GAG) mutations in the CCC phage and nearly completely prevents generation by UV of Su+ (GAG) mutations in the GGG*

*phage, in which UV-induced pyrimidine photo-products can be **formed** only in the vicinity of the target triplet.*
b) *These results indicate that OmpR stabilizes the formation of an RNA polymerase-promoter complex, possibly a closed promoter complex, and that a transcription activator can **serve** not only as a positive but also as a negative regulator for gene expression.*

Whilst the "ability" sense is not central to the interpretation of modality, the recognition of "theoretical possibility" may be more important: stating that an event has the *potential* to happen is different from stating that it *does* (always) happen. Thus, further investigation of lexical markers of theoretical possibility will help to build upon our current categorisation model.

7.3 *Point Of View* Information

Although we suggested that there are a number of textual clues that can be used to determine the *Point of View* of a statement, our annotation experiment revealed that such explicit evidence is quite sparse, at least in abstracts. Occasionally, the sentence contains words or phrases such as *we*, *our results*, *in this study*, etc. allowing the *Point Of View* to be determined as the author(s) of the abstract. In other cases, looking at the wider surrounding context, i.e. in neighbouring sentences or even within the whole abstract, is necessary. Although our annotation assumes the lack of an explicit *Point of View* marker to indicate the *writer* point of view, further analysis of these cases must be carried out.
During annotation, however, we identified some potential additional clues that can help to determine the value of this dimension.
Consider the phrase *these results*. On its own, this provides no explicit information about the point of view of the accompanying statement. However, when occurring as the subject of *suggest* (especially in the present tense), it is normally the case that the deduction has been carried out by the author(s), as illustrated in the following example:

These results [WRITER] *suggest* [DEDUCTIVE] *that both locally and regionally targeted mutagenesis is **affected** by overproduction of the epsilon subunit.*

The *writer* value can also be assumed in such contexts when other verbs in the *deductive* and *sensory* categories are used, e.g. *indicate, imply, appear*, etc, particularly when in the present tense with an inanimate subject. An exception is when there is explicit reference to another author or work. If there is an impersonal subject, e.g. *It is suggested*, then greater contextual evidence would be required, as the point of view is ambiguous.
A further example concerns *Certainty Level* markers within the *absolute* category, which generally denote well-established facts within the community. When such a certainty level marker is present, we can assume that the statement does not correspond only to the author's personal point of view. An example is shown below:

Near the amino terminus is the sequence 35GLSGSGKS, which exemplifies a motif known [ABSOLUTE] *to **interact** with the beta-phosphoryl group of purine nucleotides.*

8. Conclusion

We have presented a scheme for classifying modality in biomedical texts according to three different dimensions, namely *Knowledge Type, Certainty Level* and *Point of View*. In many cases, textual clues can be used fairly reliably to determine the correct classification of statements according to these dimensions. The results from a preliminary annotation experiment based on this scheme confirm this hypothesis.
Contextual information surrounding modal lexical items can also be important in determining the correct modal value of statements. Shallow parsing (i.e. chunking), on the top of which event annotation and modality annotation are carried out, can help to identify such information. This is in agreement with Medlock & Briscoe (2006), who suggest that linguistically-motivated knowledge may help to boost the performance of an automatic hedge classification system.
Our preliminary results suggest that many modal items in our list are fairy stable semantically when modifying bio-events. However, the correct interpretation of modal auxiliaries within the domain is more problematic, and is thus an area for further research. Our experiment also served to highlight certain weaknesses in the original model, e.g. the lack of a *Knowledge Type* category corresponding to reported facts. A further potential weakness in our results is that, whilst examples supporting all of our proposed categories were found, there is a strong bias towards certain categories. This may be because our preliminary study was based only on abstracts.
In the future, we plan to carry out further experiments to reinforce the validity of our proposed classification. These include involving multiple annotators (including biologists) to provide inter-annotator agreement statistics, as well as applying our scheme to full texts, where we can expect a greater variability of modal expression to be encountered.

9. Acknowledgements

The work described in this paper has been funded by the European BOOTStrep project (FP6 - 028099). We would also like to thank Philip Cotter for his generous help with biomedical issues.

10. References

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M. & Sherlock, G. (2000). Gene Ontology: tool for the unification of biology. *Nature Genetetics,* 25, pp 25--29.

Coates, J. (1995). The expression of root and epistemic possibility in English. In B. Aarts & C. F. Meyer (Eds.), *The Verb in Contemporary English. Theory and Description.* Cambridge: Cambridge University Press, pp 145--156.

Collins, P. C. (2006). Can and may: monosemy or polysemy?. In I. Mushin & M. Laughren, (Eds.), *Annual Meeting of the Australian Linguistic Society*, Brisbane, Australia.

DiMarco, C., & Mercer, R.E. (2004). Hedging in

scientific articles as a means of classifying citations. In *Working Notes of the American Association for Artificial Intelligence (AAAI) Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*, pp 50--54.

Dixon, R. M. W. (2005) *A Semantic Approach to English Grammar*. Oxford: Oxford University Press.

Hahn, U. & Wermter, J. (2006). Levels of Natural Language Processing for Text Mining. In S. Anananiadou & J. McNaught (Eds.), *Text Mining for Biology and Biomedicine.* London: Artech House, pp. 13--42.

Holmes, J. (1988). Doubt and certainty in ESL textbooks. *Applied Linguistics,* 13(2), Oxford: Oxford University Press, pp. 21--44.

Hoye, L. (1997). *Adverbs and Modality in English*. London, New York: Longman.

Hyland, K. (1996a). Talking to the Academy: Forms of Hedging in Science Research Articles. *Written Communication*, 13(2), pp.251--281.

Hyland, K. (1996b). Writing Without Conviction? Hedging in Science Research Articles. *Applied Linguistics* 17(4), Oxford: Oxford University Press, pp. 433--454.

Light, M., Qiu, X.Y. & Srinivasan, P. (2004). The Language of Bioscience: Facts, Speculations, and Statements In Between. In *Proceedings of BioLink 2004 Workshop on Linking Biological Literature, Ontologies and Databases: Tools for Users*, pp.17--24.

McNaught, J & Black, W. (2006). Information Extraction. In S. Ananiadou & J. McNaught (Eds), *Text Mining for Biology and Biomedicine.* London: Artech House, pp. 143--178.

Medlock, B. & Briscoe, T. (2007). Weakly Supervised Learning for Hedge Classification in Scientific Literature. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Prague, Czech Republic, pp. 992--999.

Morton T. & LaCivita J. (2003). Word-Freak: an open tool for linguistic annotation. In *Proceedings of HLT/NAACL-2003*, pp. 17--18.

Palmer, F. (1986). *Mood and modality.* Cambridge: Cambridge University Press

Rizomilioti, V. (2006). Exploring Epistemic Modality in Academic Discourse Using Corpora. *Information Technology in Languages for Specific Purposes* (7), pp. 53--71.

Rubin, V. (2007). Stating with Certainty or Stating with Doubt: Intercoder Reliability Results for Manual Annotation of Epistemically Modalized Statements. In *Proceedings of The Human Language Technologies Conference: The Annual Conference of the North American Chapter of the Association for Computational Linguistics*, *Companion Volume*, pp. 141--144.

Rubin, V. L., Liddy, E. D., & Kando, N. (2005). Certainty Identification in Texts: Categorization Model and Manual Tagging Results. In J. G. Shanahan, Y. Qu & J. Wiebe (Eds.), *Computing Attitude and Affect in Text: Theory and Applications (the Information Retrieval Series*. New York: Springer-Verlag, pp. 61--76.

Sauri, R., Verhagen, M., & Pustejovsky, J. (2006). Annotating and Recognizing Event Modality in Text. In *Proceedings of the 19th International FLAIRS Conference, FLAIRS 2006*. Melbourne Beach, Florida, pp. 333--339.

Thompson, P., Cotter, P., Ananiadou, S., McNaught, J., Montemagni, S., Trabucco, A., Venturi, G., (2008). Building a Bio-Event Annotated Corpus fro the Acquisition of Semantic Frames from Biomedical Corpora, to appear in *Proceedings of Sixth International Conference on Language Resource and Evaluation (LREC 2008)*.

Tsuruoka, Y., Tateishi, Y., Kim, J-D., Ohta, T., McNaught, J., Ananiadou, S. & Tsujii, J. (2005). Developing a Robust Part-of-Speech Tagger for Biomedical Text, In *Advances in Informatics - 10th Panhellenic Conference on Informatics*, pp 382--392.

Wilbur, W.J., Rzhetsky, A. and Shatkay, H. (2006) New directions in biomedical text annotation: definitions, guidelines and corpus construction. *BMC Bioinformatics* 7:356

Appendix A: Lexical Modality Markers

Knowledge Type Markers

Speculative **-** *assume, assumption, belief, believe, claim, conceivable, estimate, expect, expectation, hypothesise, hypothesis, hypothetical, in principle, in theory, judge, model, notion, predict, prediction, proposal, propose, speculate, suggest*[3]*, suggestion, suppose, suspect, theory, think, to our knowledge, view.*

Deductive – *argue, argument, deduce, imply, indicate, indication, infer, interpret, interpretation, suggest*[4]*.*

Demonstrative - *conclude, conclusion, confirm, confirmation, demonstrate, find, finding, proof, prove, report, reveal, show.*

Sensory - *apparent, apparently, appear, observation, observe, evidence, evident, seem, see.*

Certainty markers

Absolute - *certainly, known.*

High - *consistent with*[5]*, clear, clearly, generally, in agreement with*[5]*, likelihood, likely, normally, obviously, probability, probable, probably, strongly*[6]*, support*[5]*, would.*

Medium – *can, could, feasible, may, might, perhaps, possibility, possible, potential, potentially*.

Low – *unlikely, unknown.*

[3] with a human subject, e.g. *We suggest that …* or in the passive voice, e.g. *It is suggested that…*

[4] with an inanimate subject, e.g. *The results suggest that …*

[5] Often used to lower the speculation (and hence increase the certainty) of a speculative statement, e.g. *These results are consistent with the view that …*

[6] Often used to strengthen the certainty of deductive or speculative propositions, *e.g. The results strongly suggest that …*