



Categorising Modality in Biomedical Texts

Paul Thompson¹, Giulia Venturi², John McNaught¹,
Simonetta Montemagni² and Sophia Ananiadou¹

¹National Centre for Text Mining, University of Manchester

²Istituto di Linguistica Computazionale, CNR, Italy

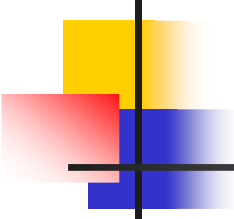
Workshop on **Building and Evaluating Resources for
Biomedical Text Mining**

LREC 2008 - 26th May 2008, Marrakech



The starting point

- Text processing systems generally focus on **factual** language
- The (potentially) **non-factual** nature of propositions can be indicated in several ways, e.g.
 - author's level of certainty towards a statement
 - whether the statement is a speculation or based on facts/evidence
- Expression of such information is called *epistemic modality*
 - Vital to take into account to correctly interpret texts



Expressing epistemic modality in biological texts

- Epistemic modality commonly conveyed in texts through:
 - discourse-based strategies
 - ***We do not know whether*** *the increase in intensity from 250 ...*
 - lexical items (i.e. words and phrases)
 - *EvgA is **likely** to directly upregulate operons ...*
 - *We **speculate** that the product of this gene is involved in ...*
- Speculations are realised **lexically** in 85% of cases (see Hyland, 1996a; 1996b)

Correct interpretation of statements in biomedical texts depends on the accurate recognition and categorisation of modal words or phrases relevant within the biomedical domain



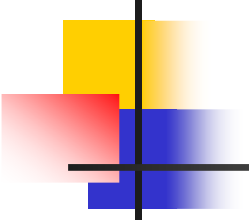
The goal

- Collect preliminary list of words and phrases expressing epistemic modality in biomedical texts
- Categorise these according to type of information expressed
- Test the categorisation through small annotation experiment



Collecting modal lexical items

- Resources used:
 - Rizomilioti (2006)
 - comprehensive list of modal lexical items from biology, archeology and literary criticism corpora
 - E. Coli corpus (provided by EBI)
 - approx 30,000 MEDLINE abstracts on subject of E. Coli
- Refine Rizomilioti's list in 2 ways:
 - Examine subset of E. Coli corpus to identify further relevant modal items
 - Eliminate items without modal sense in the corpus
 - Low occurrence thresholds (< 10)
 - Examination of usage in context



Categorising modal lexical items: what to account for?

- whether speculative or based on factual data (e.g. experimental findings)
 - *We **believe** that the inhibitor interacts ..*
 - *We have **shown** that the open reading frame encodes ...*
- the type or source of the information
 - ***Trifonov [38]** has suggested that the 530 loop ...*
 - ***Our data** demonstrate that ICG-001 has no effect ...*
- interpretation of evidence
 - *These changes **appeared** to involve both assimilatory ...*
 - *Our results **imply** that the aromatic ring to which ...*
- level of certainty
 - *The DNA-binding properties **may** indicate that ...*
 - *EvgA is **likely** to directly upregulate operons ...*



Categorising modal lexical items: a multi-dimensional model

- Different words/phrases express different information
 - ***We suggest*** that these two proteins ***may*** form a complex in the membrane....
- Therefore we propose a multi-dimensional model:
 - ***Knowledge Type***
 - the *type of knowledge* that underlies a statement, i.e. speculation or based on evidence
 - *how* the evidence is to be interpreted
 - ***Level of Certainty***
 - *how certain* the author is about the statement
 - ***Point of View***
 - whether the statement is based on the author's own point of view or experimental findings, or that of a cited work



Categorising modal lexical items: proposed scheme

- Knowledge Type examples:
 - Speculative: *predict, hypothesis, view, in theory*
 - Deductive: *interpret, indication, infer, imply*
 - Sensory: *observation, see, appear*
 - Demonstrative: *show, confirm, demonstrate*
- Certainty Level examples:
 - Absolute: *certainly, known*
 - High: *likely, probably, generally*
 - Moderate: *possibly, perhaps, may, could*
 - Low: *unlikely, unknown*
- Point of View
 - Writer: *we, our results*
 - Other: citations



Categorising modal lexical items: context of modal items

- Context can be important
 - ***We suggest that ..***
 - ***The results suggest that ..***
- Certain words/phrases only express certainty when combined with *Knowledge Type* items
 - *The results **strongly** suggest that ...*
 - *These findings are **in agreement with** the view that ...*



Testing the scheme: a small annotation experiment

- Work carried out as part of BOOTStrep project
 - Building bio-lexicon and bio-ontology
 - Creation of *gene regulation* event corpus
 - semantic arguments of verbs and nominalised denoting gene regulation events are annotated
 - facilitates extraction of semantic frames
- Annotated events are used as starting point for modality annotation
 - lexical modality markers are searched for in each sentence containing annotated gene regulation events
- 1 annotator with linguistic background
 - Supported given by 2 other researchers with biological and linguistic expertise



Testing the scheme: an example of annotation (1)

We suggest that overproduction of SlyA in

Agent

hns(+) E. coli derepresses clyA transcription

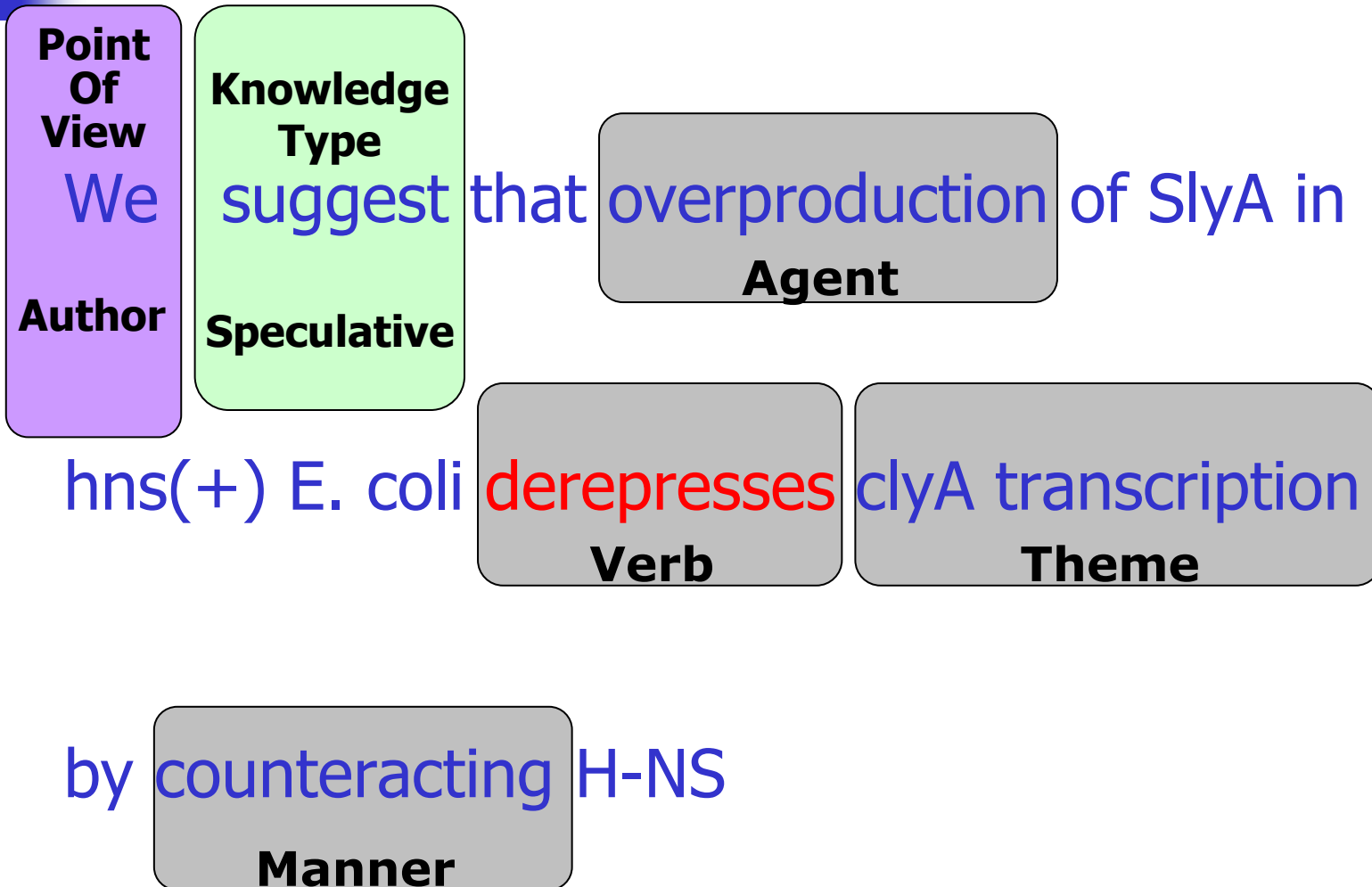
Verb

Theme

by counteracting H-NS

Manner

Testing the scheme: an example of annotation (2)



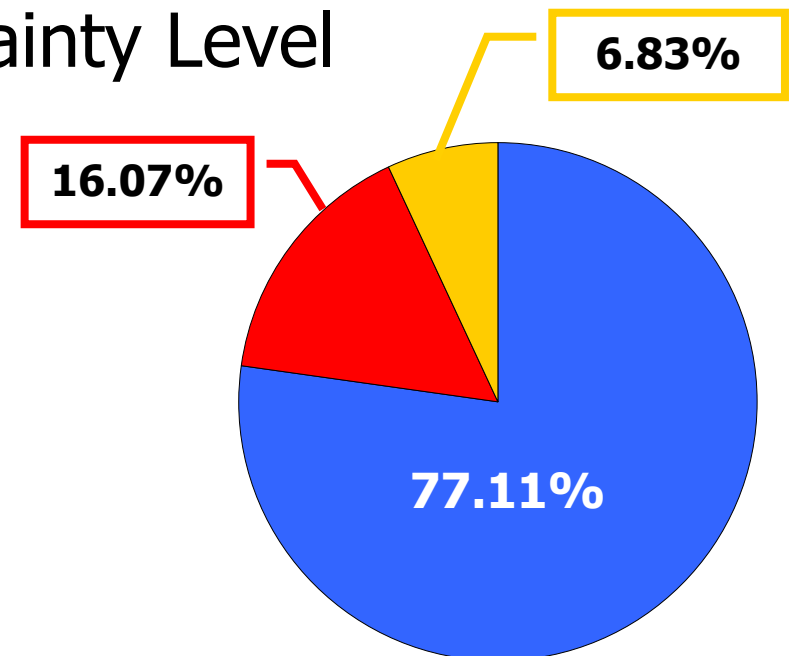
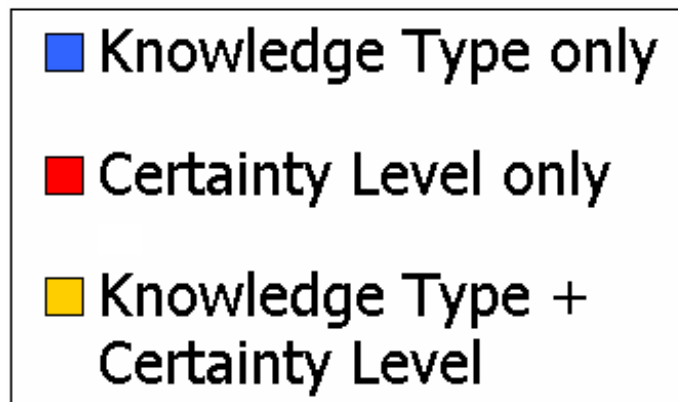


Annotation results

- 202 abstracts annotated with modality information
- Preliminary results show that modal information is reasonably sparse
 - 249 events (16.95%) out of 1469 annotated gene regulation events carry modality information
- However, allows initial verification of categorisation scheme and identification of problematic cases

Annotation results: distribution of modality marker dimensions

- Knowledge Type and Certainty Level



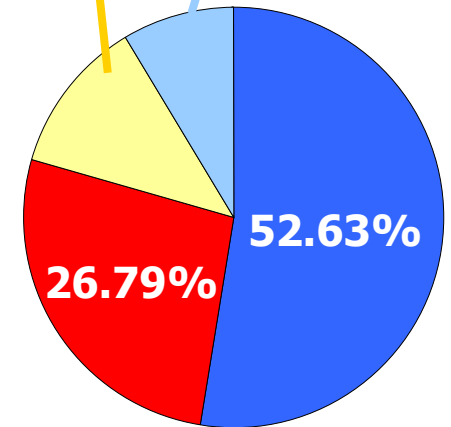
- If values were assigned to Knowledge Type and/or Certainty Level dimensions, Point of View dimension also instantiated

Knowledge Type Dimension

- Most items fairly stable semantically
 - however, *seem* has both sensory *and* speculative aspects to its meaning
- Possible need to add further value for statements that introduce facts without taking particular stance
 - e.g. *The regulator of the operon yjfS-X (ula operon) is reported to be **involved** in L-ascorbate metabolism*
- Propose use of a *quotative* category

11.96%

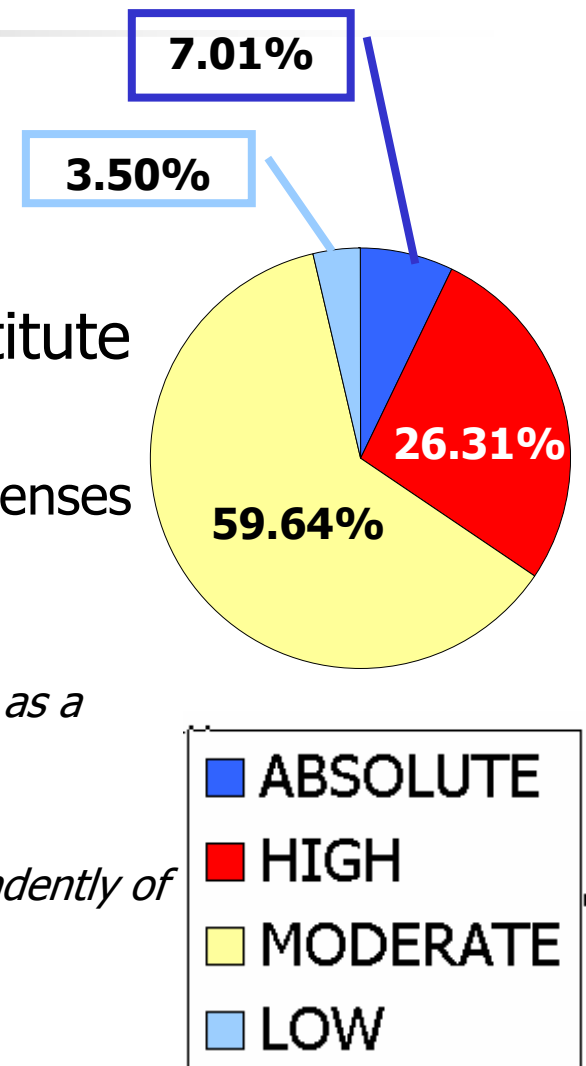
8.61%



■ DEMONSTRATIVE
■ DEDUCTIVE
■ SENSORY
■ SPECULATIVE

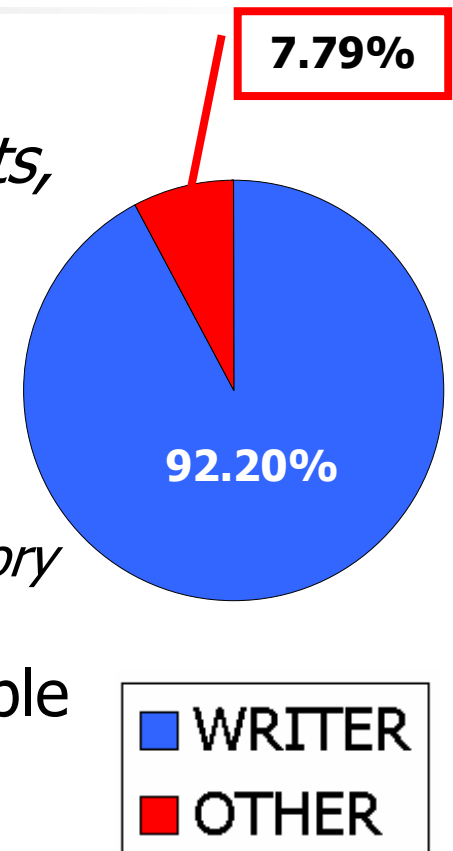
Certainty Level Dimension

- Many fairly stable semantically
 - adjectives and adverbs such as *probable*, *possibly* or *likely*
- Modal auxiliaries (e.g. *can*, *may*) constitute 40.35% of certainty level markers
 - problematic interpretation due to multiple senses
 - *Moderate* level of certainty
 - Theoretical probability (potential to occur)
 - *A transcription activator **can** serve not only as a positive but also ...*
 - Ability
 - *the promoter **may** be **transcribed** independently of the *nrdA* gene.*
 - Permission



Point of View dimension

- Quite sparse evidence
 - Occasional presence of *we, our results, in this study* etc.
- Other clues may be used
 - e.g. "These results *suggest* that ..."
normally indicates author deduction
 - Also applies to other *deductive* and *sensory* verbs e.g. *indicate, imply, appear* etc.
- If no explicit lexical evidence was available for this dimension, a "default" value of *writer* was assigned
 - i.e. it was assumed that the Point of View was expressed implicitly





Conclusions (1)

- Preliminary experimental results suggest that textual clues can classify modal statements along multiple dimensions
 - *Knowledge Type, Certainty Level & Point of View*
- Contextual information can also be important
 - Shallow parsing can help to identify this
- Apparent semantic stability amongst lexical items when modifying biological events
 - Modal auxiliaries more problematic - further research required



Conclusions (2)

- Certain problems identified
 - Lack of *Knowledge Type* category corresponding to reported facts
 - Strong bias towards certain categories within dimensions
 - May be due to dealing only with abstracts
- Future work
 - Further experiments with multiple annotators, including biologists
 - Will allow inter-annotator agreement scores to be calculated
 - Annotation of full texts
 - Wider range of modal expression is expected