A Three-Way Perspective on Scientific Discourse Annotation for Knowledge Extraction

Maria Liakata
Aberystwyth University, UK / EMBL-EBI, UK
liakata@ebi.ac.uk

Paul Thompson
University of Manchester, UK
paul.thompson@manchester.ac.uk

Anita de Waard
Elsevier Labs, USA / UiL-OTS, Universiteit Utrecht, NL
a.dewaard@elsevier.com

Raheel Nawaz
University of Manchester, UK
raheel.nawaz@cs.man.ac.uk

Henk Pander Maat
UiL-OTS, Universiteit Utrecht, NL
h.l.w.pandermaat@uu.nl

Sophia Ananiadou
University of Manchester, UK
sophia.ananiadou@manchester.ac.uk

Abstract

This paper presents a three-way perspective on the annotation of discourse in scientific literature. We use three different schemes, each of which focusses on different aspects of discourse in scientific articles, to annotate a corpus of three full-text papers, and compare the results. One scheme seeks to identify the core components of scientific investigations at the sentence level, a second annotates meta-knowledge pertaining to bio-events and a third considers how epistemic knowledge is conveyed at the clause level. We present our analysis of the comparison, and a discussion of the contributions of each scheme.

1 Introduction

The literature boom in the life sciences over the past few years has sparked increasing interest into text mining tools, which facilitate the automatic extraction of useful knowledge from text (Ananiadou et al., 2006; Ananiadou & McNaught, 2006; Zweigenbaum et al., 2007; Cohen & Hunter, 2008). Most of these tools have focussed on entity recognition and relation extraction and with few exceptions, e.g., (Hyland, 1996; Light et al., 2004; Sándor, 2007; Vincze et al., 2008), do not take into account the discourse context of the knowledge extracted. However, failure to take this context into account results in the loss of information vital for the correct interpretation of extracted knowledge, e.g. the scope of the relations, or the level of certainty with which they are expressed. A particular piece of knowledge may represent, e.g., an accepted fact, hypothesis, results of an experiment, analysis based on experimental results, factual or speculative statements etc. Furthermore, this knowledge may represent the author's current work, or work reported elsewhere. The ability to recognise different discourse elements automatically provides crucial information for the correct interpretation of extracted knowledge, allowing scientific claims to be linked to experimental evidence, or newly reported experimental knowledge to be isolated. The importance of categorising such knowledge becomes more pronounced as analysis moves from abstracts to full papers, where the content is richer and linguistic constructions are more complex (Cohen et al., 2010). Analysis of full papers is extremely important, since less than 8% of scientific claims occur in abstracts (Blake, 2010).

Various different schemes for annotating discourse elements in scientific texts have been proposed. The schemes vary along several axes, including perspective, motivation, complexity and the granularity of the units of text to which the scheme is applied. Faced with such variety, it is important to be able to select the best scheme(s) for the purpose at hand. Answers to questions such as the following can help in the selection process:

1. What are the relative merits of the different schemes?
2. What are the similarities and differences between schemes?
3. Can annotation according to multiple schemes provide enhanced information?

Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 37–46,
Jeju, Republic of Korea, 12 July 2012.

Category	Description
Hypothesis	An unconfirmed statement which is a stepping stone of the investigation
Motivation	The reasons behind an investigation
Background	Generally accepted background knowledge and previous work
Goal	A target state of the investigation where intended discoveries are made
Object-New	An entity which is a product or main theme of the investigation
Object-New-Advantage	Advantage of an object
Object-New-Disadvantage	Disadvantage of an object
Method-New	Means by which authors seek to achieve a goal of the investigation
Method-New-Advantage	Advantage of a Method
Method-New-Disadvantage	Disadvantage of a Method
Method-Old	A method mentioned pertaining to previous work
Method-Old-Advantage	Advantage of a Method
Method-Old-Disadvantage	Disadvantage of a Method
Experiment	An experimental method
Model	A statement about a theoretical model or framework
Observation	The data/phenomena recorded in an investigation
Result	Factual statements about the outputs, interpretation of observations
Conclusion	Statements inferred from observations & results

Table 1. The CoreSC Annotation scheme: layers 1 & 2

4. Is there any advantage in merging annotation schemes or is it better to allow complementary and different dimensions of scientific discourse annotation?

As a starting point to addressing such questions, we provide a comparison of three different schemes for the annotation of discourse elements within scientific papers. Each scheme has a different perspective and motivation:, one is content-driven, seeking to identify the main components of a scientific investigation, another is driven by the need to describe events of biomedical relevance and the third focusses on how epistemic knowledge is conveyed in discourse.

These different viewpoints mean that the schemes vary in both the type and complexity of the discourse elements identified, as well as the types of units to which the annotation is applied, i.e. complete sentences, segments of sentences, or specific relations/events occurring within these sentences. To facilitate the comparison, we have annotated three full papers according to each of the schemes. The analysis resulting from this three-way annotation considers mappings between schemes, their relative merits, and how the information annotated by the different schemes can complement each other to provide enriched details about knowledge extracted from the texts.

In the following sections, we firstly provide a description of the three schemes, and then explain how they have been used in our corpus annotation. Finally we discuss the results from the comparison, and the features of each scheme.

2 Sentence annotation: CoreSC scheme

The reasoning behind this scheme is that a paper is the human-readable representation of a scientific investigation. Therefore, the goal of the annotation is to retrieve the content model of scientific investigations as reflected within scientific discourse. The hypothesis is that there is a set of core scientific concepts (CoreSC), which constitute the key components of a scientific investigation. CoreSCs consist of 11 concepts originating from the CISP (Core Information about Scientific Papers) meta-data (Soldatova & Liakata, 2007), which are a subset of classes from the EXPO ontology for the description of scientific experiments (Soldatova & King, 2006). The CoreSCs are: *Motivation, Goal, Object, Background, Hypothesis, Method, Model, Experiment, Observation, Result* and *Conclusion*.

The CoreSC scheme (Liakata et al., 2010; Liakata et al., 2012) implements the above-mentioned concepts as a 3-layered sentence-based annotation scheme. This means that each sentence in a document is assigned one of the 11 CoreSC concepts. The scheme also considers a layer designated to properties of the concepts (e.g. New Method vs Old Method) as well as identifiers which link instances of the same concept across sentences. A short definition of CoreSC categories and their properties can be found in Table 1.

The CoreSC scheme is accompanied by 47-page annotation guidelines, and has been used by 16 domain experts to annotate a corpus of 265 full papers from physical chemistry & biochemistry (Liakata & Soldatova, 2009; Liakata et al., 2010). This corpus consists of 40,000 sentences, containing over 1 million words and was developed in three phases (for details see Liakata et al. (2012)). Inter-annotator agreement between experts was measured in terms of Cohen's kappa (Cohen, 1960) on 41 papers and ranged between 0.5 and 0.7. Machine learning classifiers have been trained on the CoreSC corpus, achieving > 51% accuracy across the eleven categories. The most accurately predicted category is *Experiment,* the category describing experimental methods (Liakata et al., 2012). Classifiers trained on 1000 Biology abstracts annotated with CoreSC have obtained an accuracy of over 80% (Guo et al., 2010). Models trained on the CoreSC corpus papers have been used to create automatic summaries of the papers, which have been evaluated in a question answering task (Liakata et al., 2012). Lastly, the CoreSC scheme was used to annotate 50 papers from Pubmed Central pertaining to Cancer Risk Assessment. A web tool (SAPIENTA[1]) allows users to annotate their full papers with Core Scientific concepts, and can be combined with manual annotation. A UIMA framework [2] implementation of this code for large-scale annotation of CoreSC concepts is in progress.

3 Event annotation: Meta-knowledge for bio-events

The motivation for this annotation scheme is to allow the training of more sophisticated event-based information extraction systems. In contrast to the sentence-based scheme described in section 2, this scheme is applied at the level of *events* (Ananiadou et al., 2010), of which there may be several within a single sentence.

3.1 Bio-Events

Events are template-like, structured representations of pieces of knowledge contained within sentences. Normally, events are "anchored" to a *trigger* (typically a verb or noun) around which the knowledge expressed is organised. Each event has one of more participants, which describe different aspects of the event. Participants can correspond to entities or other events, and are often labelled with semantic roles, e.g., CAUSE, THEME, LOCATION, etc. The work described here focusses specifically on bio-events, which are complex structured relations representing fine-grained relations between bio-entities and their modifiers. Figure 1 provides some examples of bio-events. Event extraction systems (Björne et al., 2009; Miwa et al., 2010; Miwa et al., 2012; Quirk et al., 2011) are typically trained on text corpora, in which events and their participants have been manually annotated by domain experts. Research into bio-event extraction has been boosted by the two recent shared tasks at BioNLP 2009/2011 (Kim et al., 2011; Pyysalo et al., In Press). Several gold standard event annotated corpora exist; examples include the GENIA Event Corpus (Kim et al., 2008), GREC (Thompson et al., 2009) and BioInfer (Pyysalo et al., 2007), in addition to the corpora produced for the shared tasks.

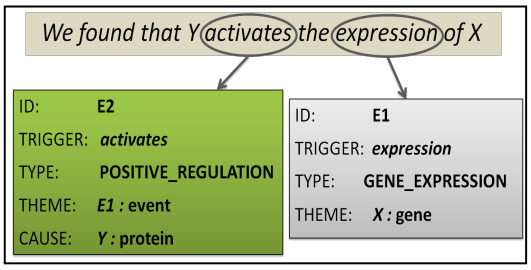

Figure 1. Bio-Event Representation

3.2 Meta-knowledge Annotation

Until recently, the only attempts to recognise information relating to the correct interpretation of events were restricted to sparse details regarding negation and speculation (Kim et al., 2011).

[1] http://www.sapientaproject.com/software
[2] http://uima.apache.org/

In order to address this problem, a multi-dimensional annotation scheme especially tailored to bio-events was developed (Nawaz et al., 2010; Thompson et al., 2011). The scheme identifies and categorises several different types of contextual details regarding events (termed *meta-knowledge*), including discourse information. Different types of meta-knowledge are encoded through five distinct dimensions (Figure 2). The advantage of using multiple dimensions is that the interplay between the assigned values in each dimension can reveal both subtle and substantial differences in the types of meta-knowledge expressed.

In the majority of cases, meta-knowledge is expressed through the presence of particular "clue" words or phrases, although other features can also come into play, such as the tense of the event trigger, or the relative position within the text.

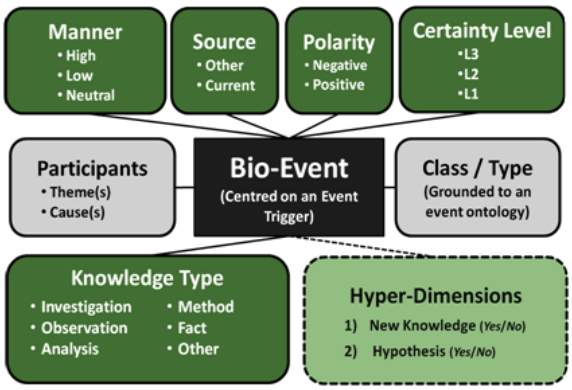

Figure 2: Meta-knowledge annotation

The annotation task consists of assigning an appropriate value from a fixed set for each dimension, as well as marking the textual evidence for this assignment. The five meta-knowledge dimensions and their values are as follows:

Knowledge Type (KT): Captures the general information content of the event. Each event is classified as one of: *Investigation* (enquiries and examinations, etc.), *Observation* (direct experimental observations), *Analysis* (inferences, interpretations and conjectures, etc.), *Fact* (known facts), *Method* (methods) or *Other* (general events that provide incomplete information or do not fit into any other category).

Certainty Level (CL): Encodes the confidence or certainty level ascribed to the event in the given text. The epistemic scale is partitioned into three distinct levels: *L3* (no expression of uncertainty), *L2* (high confidence or slight speculation) and *L1* (low confidence or considerable speculation).

Polarity: Identifies negated events. Negation is defined as the absence or non-existence of an entity or a process.

Manner: Captures information about the rate, level, strength or intensity of the event, using three values: *High*, *Low*, or *Neutral* (no indication of rate/intensity).

Source: Encodes the source of the knowledge being expressed by the event as *Current* (the current study) or *Other* (any other source).

Of these five dimensions, only *KT, CL* and *Source* were considered during the comparison with the other two schemes, since they are directly related to discourse analysis.

The GENIA event corpus, consisting of 1000 abstracts with 36,115 events (Kim et al., 2008) has been annotated with meta-knowledge by 2 annotators, supported by 64-page annotation guidelines [3] (Thompson et al., 2011). Inter-annotator agreement rates ranged between 0.84–0.93 (Cohen's Kappa). Research has been carried out into the automatic assignment of Manner values to events (Nawaz et al., In Press). In addition, the EventMine-MK service (Miwa et al., In Press), based on EventMine (Miwa et al., 2010) facilitates automatic extraction of biomedical events with meta-knowledge assigned. The performance of EventMine-MK in assigning different meta-knowledge values to events ranges between 57% and 87% (macro-averaged F-Score) on the BioNLP'09 Shared Task corpus (Kim et al, 2011). EventMine-MK is available as a component of the U-Compare interoperable text mining system[4] (Kano et al., 2011).

4 Clause annotation: Segments for epistemic knowledge

The third scheme we consider uses a Discourse Segment Type classification of segments at, roughly, a clause level, i.e., each segment has a main verb. This means that the level of granularity of argumentational elements in this scheme lies between the other two schemes, i.e. it is usually more granular than CoreSC, but sometimes less granular than the event-based scheme.

[3] http://www.nactem.ac.uk/meta-knowledge/
[4] http://www.nactem.ac.uk/ucompare/

Segment	Description	Examples
Fact	knowledge accepted to be true, a known fact.	*mature miR-373 is a homolog of miR-372,*
Hypothesis	a proposed idea, not supported by evidence	*This could for instance be a result of high mdm2 levels*
Problem	unresolved, contradictory, or unclear issue	*However, further investigation is required to demonstrate the exact mechanism of LATS2 action*
Goal	research goal	*To identify novel functions of miRNAs,*
Method	experimental method	*Using fluorescence microscopy and luciferase assays,*
Result	a restatement of the outcome of an experiment	*all constructs yielded high expression levels of mature miRNAs*
Implication	an interpretation of the results, in light of data	*our procedure is sensitive enough to detect mild growth differences*
Other-Hypothesis	an idea proposed by others	*[It is generally believed that] transcription factors are the final common pathway driving differentiation]*
Regulatory-Hypothesis	a matrix clause introducing a hypothesis	*It is generally believed that [transcription factors are the final common pathway driving differentiation]*

Table 2: Discourse Segment Types

The segment annotation scheme identifies a taxonomy of discourse segment types that seem to be exclusive and useful (de Waard & Pander Maat, 2009). Three classes of segment types are defined:

- Basic segment types: segments referring directly to the topic of study – see Table 2.
- 'Other'-segment types: segments referring to conceptual or experimental work in other research papers than the current one
- Regulatory segment types: 'regulatory' clauses that control and introduce other segments.

A list of segment types is presented in Table 2; further details, including a list of all segment types and correlations with verb tense can be found in de Waard & Pander Maat (2009). The focus of this work is to identify linguistic features that characterise these discourse segment types, according to three aspects:

- Verb tense, aspect, mood and voice
- Semantic verb class
- Epistemic modality markers

So far, 6 full-text papers (comprising about 2300 segments) have been manually annotated with segment types and correlated with the above features. A first automated validation was promising (de Waard, Buitelaar and Eigener, 2009). The need for parsing at a clause level is especially prominent in biological text, since specific semantic roles are played by particular clause types. We give four examples of typical clause constructions that play a specific rhetorical role: firstly, reporting clauses are often sentence-initial 'that' matrix clauses (1a):

1. a. This suggests that
1.b. miR-372 and miR-373 caused the observed selective growth advantage.

Secondly, descriptions confirming certain accepted characteristics of biological entities are often given as nonrestrictive relative clauses (2b):

2.a. We also generated BJ/ET cells expressing the RASV12-ERTAM chimera gene,
2. b. which is only active when tamoxifen is added

Thirdly, a subordinate gerund clause is often used to describe a method (3a), with a main (finite) clause describing a result (3b) and fourthly, experimental goals are often given as a (mostly sentence-initial) clause with a to-infinitive (4a) often preceding a past-tense methods clause (4b).

3. a. Using fluorescence microscopy and luciferase assays,
b. we observed potent and specific miRNA activity expressed from each miR-Vec (Figure S2).
4. a. To identify miRNAs that can interfere with this process
4. b. we transduced BJ/ET fibroblasts with miR-Lib

However, the lack of simple robust clause parsers has prevented the automated identification of semantic roles at the clause level. Therefore, this scheme has so far only been manually

implemented. Despite being less widely implemented than the other two schemes, we believe that the segment scheme offers some useful pointers for linguistic features that can identify particular rhetorical classes in the text, and secondly, offers an interesting perspective on the fact that in biological text, several rhetorical moves are made within a single sentence.

5 Data and methods

Three papers already annotated according to the GENIA event annotation scheme (Kim et al., 2008), were further annotated according to the three annotation schemes described above. We obtained all corresponding CoreSCs, events and segments per sentence. Each sentence has a single CoreSC annotation and one or more segment annotations (depending on the number of clauses). Event annotations in a sentence may range from zero to multiple, according to whether any relevant biomedical events are described in the sentence.

Events within a sentence are mapped to segments by identifying which segment contains the trigger for a particular event. The three meta-knowledge dimensions for events considered in this comparison, i.e., KT, CL and Source, result in 16 different combinations of values encountered in the three papers. The numbers for CoreSC and Segment labels encountered were 12 and 22, respectively. Confusion matrices were obtained for each paper and for each pair of annotation schemes. Note that, as bio-events are largely unconcerned with describing methodology, the *Methods* sections of these papers do not contain event annotation or meta-knowledge annotation. The pairwise confusion matrices from each paper were combined, resulting in three matrices (Tables 3, 4 and 5), which describe the associations between the annotation schemes in the three papers examined. We have highlighted the highest frequencies per row and where appropriate also the highest values per column. The use of two different colours aims to facilitate readability.

6 Results and Discussion

We present the results from analysing the pairwise confusion matrices for the three schemes and discuss the merits of each scheme.

6.1 Event Meta-knowledge v. CoreSC

In Tables 3 (and 5), the meta-knowledge categories combine KT, CL and Source ((O)ther) values. Table 3 shows some straightforward and expected mappings, e.g.,Method (Met,L3) events are almost always found within CoreSC Experiment or Method sentences, whilst Investigation events (Inv,L3) occur most frequently within CoreSC Goal or Motivation sentences.

For other categories, information from the two schemes can complement each other in different ways. For example, KT and Source information about events can help to distinguish different types of information within CoreSC Background sentences (top left corner of Table 3). Such information mainly corresponds to facts, observations from previous studies, or analyses of information. Conversely, information from the CoreSC scheme can help to further classify the interpretation of events. For example, events with an analytical interpretation (Ana,L1,L2,L3) may occur as background information to a study (Bac), as hypotheses (Hyp), as part of observations (Obs), when reporting the results of the current study (Res) or when making concluding remarks about the study (Con). CoreSCs can also help to further refine events relating to outcomes (Obs,L3) according to whether they pertain to (Obs)ervations, (Res)ults or (Con)clusions.

CoreSC Conclusion, Result and Observation sentences contain mainly Observation events concerned with the current study. However, such sentences often also include an analytical part, with varying levels of certainty, which event information can help to isolate. The CL annotated for events is also useful in helping to determine the confidence with which information is stated in CoreSC Conclusion and Hypothesis sentences.

Due to the nature of bio-event annotation, only a small number of events correspond to methods. Thus, CoreSC provides a more detailed characterisation of method-related sentences, i.e., Experiment, Method_New, Model and Object.

6.2 Discourse Segments v. CoreSC

In most cases, there seems to be natural mapping between the two schemes (See Table 4). CoreSC Observation maps to *Result*, CoreSC Method and Experiment map to *Method*, CoreSC Hypothesis maps to *Hypothesis*, CoreSC Goal maps to *Goal*,

CoreSC Conclusion maps to *Implication* and *Hypothesis,* CoreSC Result maps to *Implication* and *Result*, and *Problem* is equivalent to CoreSC Motivation. The bulk of CoreSC Background maps to *Fact* and *Other-Implication*, but the "Other" Segment categories provide a substantial refinement of the CoreSC Background category.

	Bac	Con	Exp	Goa	Hyp	Met_New	Met_Old	Mod	Mot	Obj_New	Obs	Res
0	42	24	49	7	7	25	1	13	6	7	47	54
Obs,L3,O	166	0	0	0	0	0	3	0	12	0	0	2
Ana,L3,O	33	1	0	0	0	0	0	1	0	0	0	0
Ana,L2,O	3	0	0	0	0	0	0	0	0	0	0	0
Fact,L3,O	7	0	0	0	0	0	0	0	0	0	0	0
Fact,L3	24	1	0	1	0	0	0	0	5	3	0	2
Oth,L3	125	30	0	8	16	5	3	2	8	3	9	42
Ana,L1	2	10	0	0	6	0	0	0	1	0	0	6
Ana,L2	30	15	0	1	14	0	0	2	1	0	8	33
Ana,L3	11	11	0	0	2	1	2	0	3	0	14	28
Met,L3	4	1	15	1	0	5	0	0	0	0	2	6
Inv,L2	0	0	0	0	0	0	0	0	0	0	1	1
Inv,L3	5	3	1	6	2	4	3	0	8	0	1	1
Inv,L3,O	0	0	0	0	2	0	1	0	0	0	0	0
Obs,L1	1	0	0	0	0	0	0	0	0	0	0	0
Obs,L2	1	0	0	0	0	0	0	0	0	0	1	0
Obs,L3	31	34	3	1	10	3	0	2	7	1	59	87

Table 3. Event Meta-knowledge vs CoreSC

On the other hand, CoreSC refines *Method*, *Result* and *Implication* segments. CoreSC Result may include both *Fact* and *Method* clauses, which can be captured by the Segment scheme, since annotation is performed at the clause level. CoreSC Conclusion maps to both *Implication* and *Hypothesis* segments, suggesting that there may be differences in the certainty levels of these conclusions. This is supported by preliminary classification experiments (paper in progress).

6.3 Discourse Segments v. Event Meta-Knowledge

Some straightforward mappings exist between segment and event meta-knowledge categories (Table 5). For example, Investigation events (Inv, L3) are generally found within *Goal* and *Problem* segments; Method events (Met,L3) are normally found within *Method* segments, Observation events (Obs,L3) are found mainly within *Result*, *Fact* and *Implication* segments and (Ana,L1,L2) events correspond mainly to Hypotheses and Implications.

Whilst these are similar findings to the comparison between event meta-knowledge and CoreSCs, the variance of the distribution is often smaller when mapping from Events to Segments. This is to be expected – the information encoded by many events has the scope of roughly a clause, which corresponds closely to the scope of discourse segments. This could permit cleaner one-to-one mappings between categories.

	Bac	Con	Exp	Goa	Hyp	Met_New	Met_Old	Mod	Mot	Obj_New	Obs	Res
Fact	118	3	0	3	7	0	0	1	15	7	5	34
OtherFact	70	4	0	0	0	0	0	3	1	0	0	0
OtherGoal	2	1	0	0	0	0	0	0	0	0	0	0
OtherHypothesis	14	0	0	0	0	0	0	0	0	0	0	0
OtherImplication	124	1	0	0	3	0	0	1	5	0	0	1
OtherMethod	5	0	0	0	0	0	3	0	0	0	0	2
OtherProblem	1	0	0	0	0	0	0	0	0	0	0	0
OtherResult	64	1	0	0	0	0	6	0	0	3	0	9
RegFact	1	3	0	0	0	0	0	0	0	0	0	2
Implication	13	58	0	0	2	0	0	3	1	0	3	80
RegImplication	5	6	0	1	0	0	0	0	0	0	1	10
Method	6	2	54	2	2	32	0	6	1	0	8	13
Goal	2	0	5	12	6	9	2	2	4	0	0	5
RegGoal	0	1	0	0	0	0	0	0	0	0	0	0
Hypothesis	24	31	0	5	34	1	0	5	0	0	0	12
RegHypothesis	6	4	0	0	2	0	0	1	0	0	0	2
Problem	7	6	0	0	0	0	2	0	11	0	0	2
RegProblem	0	3	0	0	0	0	0	0	0	0	0	0
Result	13	6	1	1	2	0	0	2	8	0	112	75
RegResult	1	0	0	0	0	0	0	0	0	1	1	2
Intertextual	4	0	7	0	1	0	0	0	0	0	0	3
Intratextual	2	0	1	0	0	0	0	2	0	0	8	4

Table 4: Segments vs CoreSC

Hypothesis and *Implication* segments mainly contain (Ana)lysis events. The differing certainty levels of events can help to refine information about the statements made within these segments. Likewise, these segment types could help to refine the nature of the analysis described by the event.

Similarly to the CoreSC scheme, the results suggest that *Result* segments could be refined by the meta-knowledge scheme to distinguish between results emerging from direct experimental observations, and those obtained through analysis of experimental observations. Another interesting result is that *Fact* segments can contain Fact, (Ana)lysis or (Obs)ervation events. This may suggest that *Fact* segments are actually a rather general category, containing a range of different information. Few events occur within the *Regulatory* segments, as these mainly introduce content-bearing segments.

The majority of *Method* segments and a significant number of the *Result* segments do not correspond to events, as none of the methods sections have been annotated with event information, for reasons explained previously.

	0	Ana L1	Ana L2	Ana L2,O	Ana L3	Ana L3,O	Fact L3	Fact L3,O	Met L3	Oth L3	Inv L2	Inv L3	Inv L3,O	Obs L1	Obs L2	Obs L3	Obs L3,O
Hypothesis	8	18	26	1	0	0	0	0	1	39	0	4	1	0	0	14	0
Implication	22	2	30	0	34	2	2	0	0	38	2	1	0	0	0	27	0
OtherHypothesis	0	0	3	1	0	0	0	0	0	9	0	1	0	0	0	0	0
OtherImplication	8	1	6	1	4	28	0	3	3	27	0	2	0	1	0	5	46
RegImplication	11	0	2	0	0	1	0	0	0	5	0	1	0	0	0	3	0
RegHypothesis	1	0	0	0	0	0	0	0	0	6	0	1	0	0	0	7	0
Fact	15	0	18	0	6	0	28	0	0	55	0	1	0	0	1	44	25
RegFact	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	5	0
OtherGoal	1	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0
OtherProblem	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
Method	80	0	1	0	0	0	0	0	23	9	0	2	0	0	0	8	3
OtherMethod	7	0	0	0	0	0	0	0	0	0	0	2	1	0	0	0	0
Goal	13	0	0	0	0	0	1	0	0	18	0	11	1	0	0	3	0
RegGoal	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Problem	9	4	0	0	2	0	0	0	0	5	0	8	0	0	0	0	0
RegProblem	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Result	51	0	14	0	20	0	0	0	6	18	0	0	0	0	1	103	7
OtherResult	11	0	1	0	0	1	0	1	0	10	0	0	0	0	0	12	47
OtherFact	4	0	1	0	0	2	5	3	0	7	0	0	0	0	0	2	54
RegResult	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Intertextual	13	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0
Intratextual	17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 5: Segments vs Event Meta-Knowledge

7 Related Work

A number of schemes for annotating scientific discourse elements at the sentence level have been proposed. Certain schemes have been aimed at abstracts, e.g., (McKnight & Srinivasan, 2003; Ruch et al., 2007; Hirohata et al., 2008; Björne et al., 2009). The work of Hirohata et al. (2009) has been integrated with the MEDIE service[5] (Miyao et al., 2006), allowing the user to query facts using conclusions, results, etc. For full papers, the most notable work has focussed on argumentative zoning (AZ) (Teufel et al., 1999; Teufel & Moens, 2002; Teufel et al., 2009; Teufel, 2010). An important aspect of AZ involves capturing the attribution of knowledge claims and citation function, and the scheme has been tested on information extraction and summarisation tasks with Computational Linguistics papers. AZ was modified for the annotation of biology papers by Mizuta et al. (2005) in order to facilitate information extraction, and more recently Teufel et al. (2009) extended the AZ scheme to better accommodate the life sciences and chemistry in particular, producing AZ-II.

Scientific discourse annotation has also targeted the retrieval of *speculative text* to help improve curation. For a recent overview see de Waard and Pander Maat (2012). Modality and negation in text have also been the focus of recent workshops (Farkas et al (2010), Morante & Sporleder (2012)). Finally, Shatkay et al (2008) define a multi-dimensional scheme, which combines several of the above-mentioned aspects.

Recent work has compared schemes to discover mappings and relative merits. Liakata et al. (2010) compared AZ-II and CoreSC on 36 papers annotated with both schemes and found that CoreSC provides finer granularity in distinguishing content categories (e.g. methods, goals and outcomes) while the strength of AZ-II lies in detecting the attribution of knowledge claims and identifying the different functions of background information. Guo et al. (2010) compared three schemes for the identification of discourse structure in scientific abstracts from cancer research assessment articles. The work showed a subsumption relation between the scheme of Hirohata et al. (2008), a cut-down version of the scheme proposed by Teufel et al. (2009) and CoreSC (1st layer), from general to specific.

8 Conclusion

We have compared three different schemes, each taking a different perspective to the annotation of scientific discourse. The comparison shows that the three schemes are complementary, with different strengths and points of focus. CoreSC offers a fine-grained characterisation of methods, outcomes and objectives. It has been used to annotate a collection of 265 full papers, and subsequently CoreSC recognition has been fully automated, creating the online SAPIENTA tool. The discourse segment annotation scheme can help to provide a finer-grained characterisation of background work, and could also help to split multi-clause CoreSC sentences into appropriate segments. Recognition of event meta-knowledge has been fully automated in the U-Compare framework, and the KT values of the scheme can help to provide a finer-grained analysis of certain segment and sentence types. The CL dimension also allows confidence values to be ascribed to the Conclusion, Result, Implication and Hypothesis categories of the other two schemes.

Future work will focus on annotating texts with several discourse perspectives to investigate the advantages of the schemes. Ideally we would like to propose a unified approach for scientific discourse annotation, but recognize that choices such as the unit of annotation are often task-oriented, and that users should be able to mix and match discourse segments as required. This said, the analysis in this paper paves the way for potential harmonisation, revealing points of union and intersection between the schemes.

Acknowledgements

This work has been supported through funding for Maria Liakata by JISC, the Leverhulme Trust and EBI-EMBL. It has also been supported by the BBSRC through grant number BB/G013160/1UK (Automated Biological Event Extraction from the Literature for Drug Discovery), the MetaNet4U project (ICT PSP Programme, Grant Agreement: No. 270893) and the JISC-funded ISHER project.

[5] http://www.nactem.ac.uk/medie/

References

Ananiadou, S., Kell, D.B. and Tsujii, J. (2006). Text mining and its potential applications in systems biology. *Trends Biotechnol*, 24(12): 571-9.

Ananiadou, S. and McNaught, J., Eds. (2006). *Text Mining for Biology and Biomedicine*. Boston / London, Artech House.

Ananiadou, S., Pyysalo, S., Tsujii, J. and Kell, D.B. (2010). Event extraction for systems biology by text mining the literature. *Trends Biotechnol*, 28(7): 381-90.

Björne, J., Heimonen, J., Ginter, F., Airola, A., Pahikkala, T. and Salakoski, T. (2009). Extracting Complex Biological Events with Rich Graph-Based Feature Sets. In *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task*, pp. 10-18.

Blake, C. (2010). Beyond genes, proteins, and abstracts: Identifying scientific claims from full-text biomedical articles. *Journal of Biomedical Informatics*, 43(2): 173-189.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20: 37-46.

Cohen, K.B. and Hunter, L. (2008). Getting started in text mining. *PLoS Comput Biol*, 4(1): e20.

Cohen, K.B., Johnson, H.L., Verspoor, K., Roeder, C. and Hunter, L.E. (2010). The structural and content aspects of abstracts versus bodies of full text journal articles are different. *BMC Bioinformatics*, 11: 492.

de Waard, A., Buitelaar, P., Eigner, T. (2009). *Identifying the epistemic value of discourse segments in biology texts*. Proceedings of the Eighth International Conference on Computational Semantics, pp. 351-354

de Waard, A. and Pander Maat, H. (2009). Categorizing Epistemic Segment Types in Biology Research Articles. In *Proceedings of the Workshop on Linguistic and Psycholinguistic Approaches to Text Structuring (LPTS 2009)*

de Waard, A. and Pander Maat, H. (2012). Knowledge Attribution in Scientific Discourse: A Taxonomy of Types and Overview of Features, In *Proceedings of the Workshop on Detecting Structure in Scholarly Discourse (DSDD)*, ACL 2012.

Farkas, R. Vincze, V., Móra, G., Csirik, J. and Szarvas, G. 2010. The CoNLL-2010 Shared Task: Learning to Detect Hedges and their Scope in Natural Language Text. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, Uppsala, Sweden. Association for Computational Linguistics, pp. 1- 12.

Guo, Y., Korhonen, A., Liakata, M., Silins, I., LiSun, L. and Stenius, U. (2010). Identifying the information structure of scientific abstracts: An investigation of three different schemes. In *Proceedings of BioNLP 2010*, pp. 99-107.

Hirohata, K., Okazaki, N., Ananiadou, S. and Ishizuka, M. (2008). Identifying Sections in Scientific Abstracts using Conditional Random Fields. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing*, pp. 381-388.

Hyland, K. (1996). Writing without conviction? Hedging in science research articles. *Applied Linguistics*, 17(4): 433-454.

Kano, Y., Miwa, M., Cohen, K.B., Hunter, L.E., Ananiadou, S. and Tsujii, J. (2011). U-Compare: A modular NLP workflow construction and evaluation system. *IBM Journal of Research and Development*, 55(3): 11:1-11:10.

Kilicoglu, H. and Bergler, S. (2008). Recognizing speculative language in biomedical research articles: a linguistically motivated perspective. *BMC Bioinformatics*, 9(Suppl 11): S10.

Kim, J.-D., Ohta, T. and Tsujii, J. (2008). Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics*, 9(10).

Kim, J.D., Ohta, T., Pyysalo, S., Kano, Y. and Tsujii, J. (2011). Extracting Bio-Molecular Events from Literature - The BioNLP'09 Shared Task. *Computational Intelligence*, 27(4): 513-540.

Liakata, M., Saha, S., Dobnik, S., Batchelor, C. and Rebholz-Schuhmann, D. (2012). Automatic recognition of conceptualisation zones in scientific articles and two life science applications. *Bioinformatics*, 28 (7).

Liakata, M. and Soldatova, L.N. (2009). The ART corpus. Technical Report. Aberystwth University.

Liakata, M., Teufel, S., Siddharthan, A. and Batchelor, C. (2010). Corpora for the conceptualisation and zoning of scientific papers. In *Proceedings of LREC*, pp. 2054-2061.

Light, M., Qiu, X.Y. and Srinivasan, P. (2004). The language of bioscience: Facts, speculations, and statements in between. In *Proceedings of the BioLink 2004 Workshop at HLT/NAACL*, pp. 17–24.

McKnight, L. and Srinivasan, P. (2003). Categorization of sentence types in medical abstracts. In *AMIA Annu Symp Proc*, pp. 440-4.

Miwa, M., Saetre, R., Kim, J.D. and Tsujii, J. (2010). Event extraction with complex event

classification using rich features. *J Bioinform Comput Biol*, 8(1): 131-46.

Miwa, M., Thompson, P. and Ananiadou, S. (2012). Boosting automatic event extraction from the literature using domain adaptation and coreference resolution. *Bioinformatics.*

Miwa, M., Thompson, P., McNaught, J, Kell, D.B and Ananiadou, S. (In Press). Extracting semantically enriched events from biomedical literature. *BMC Bioinformatics.*

Miyao, Y., Ohta, T., Masuda, K., Tsuruoka, Y., Yoshida, K., Ninomiya, T. and Tsujii, J. (2006). Semantic Retrieval for the Accurate Identification of Relational Concepts in Massive Textbases. In *Proceedings of ACL*, pp. 1017-1024.

Mizuta, Y., Korhonen, A., Mullen, T. and Collier, N. (2005). Zone Analysis in Biology Articles as a Basis for Information Extraction. *International Journal of Medical Informatics*,75(6): 468-487.

Morante R., and Sporleder C, (2012). Modality and negation: An introduction to the special issue. *Computational Linguistics,* 38(2): 1–38.

Nawaz, R., Thompson, P. and Ananiadou, S. (In Press). Identification of Manner in Bio-Events. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*.

Nawaz, R., Thompson, P., McNaught, J. and Ananiadou, S. (2010). Meta-Knowledge Annotation of Bio-Events. In *Proceedings of LREC 2010*, pp. 2498-2507.

Pyysalo, S., Ginter, F., Heimonen, J., Bjorne, J., Boberg, J., Jarvinen, J. and Salakoski, T. (2007). BioInfer: a corpus for information extraction in the biomedical domain. *BMC Bioinformatics*, 8: 50.

Pyysalo, S., Ohta, T., Rak, R., Sullivan, D., Mao, C., Wang, C., Sobral, B., Tsujii, J. and Ananiadou, S. (In Press). Overview of the ID, EPI and REL tasks of BioNLP Shared Task 2011. *BMC Bioinformatics*.

Quirk, C., Choudhury, P., Gamon, M. and Vanderwende, L. (2011). MSR-NLP Entry in BioNLP Shared Task 2011. In *Proceedings of BioNLP Shared Task 2011 Workshop*, pp. 155-163.

Ruch, P., Boyer, C., Chichester, C., Tbahriti, I., Geissbuhler, A., Fabry, P., Gobeill, J., Pillet, V., Rebholz-Schuhmann, D., Lovis, C. and Veuthey, A.L. (2007). Using argumentation to extract key sentences from biomedical abstracts. *Int J Med Inform*, 76(2-3): 195-200.

Sándor, Á. (2007). Modeling metadiscourse conveying the author's rhetorical strategy in biomedical research abstracts. *Revue Française de Linguistique Appliquée*, 200(2): 97-109.

Shatkay, H., Pan, F., Rzhetsky, A. and Wilbur, W.J. (2008). Multi-dimensional classification of biomedical text: toward automated, practical provision of high-utility text to diverse users. *Bioinformatics*, 24(18): 2086-2093.

Soldatova, L.N. and King, R.D. (2006). An ontology of scientific experiments. *Journal of the Royal Society Interface*, 3(11): 795-803.

Soldatova, L.N. and Liakata, M. (2007). An ontology methodology and cisp-the proposed core information about scientific papers., Aberystwyth University. Technical Report JISC Project Report.

Teufel, S. (2010). *The Structure of Scientific Articles: Applications to Citation Indexing and Summarization*. Stanford, CA, CSLI Publications.

Teufel, S., Carletta, J. and Moens, M. (1999). An annotation scheme for discourse-level argumentation in research articles. In *Proceedings of EACL*, pp. 110-117.

Teufel, S. and Moens, M. (2002). Summarizing scientific articles: experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4): 409-445.

Teufel, S., Siddharthan, A. and Batchelor, C. (2009). Towards discipline-independent argumentative zoning: Evidence from chemistry and computational linguistics. In *Proceedings of EMNLP 2009*, pp. 1493-1502.

Thompson, P., Iqbal, S.A., McNaught, J. and Ananiadou, S. (2009). Construction of an annotated corpus to support biomedical information extraction. *BMC Bioinformatics*, 10: 349.

Thompson, P., Nawaz, R., McNaught, J. and Ananiadou, S. (2011). Enriching a biomedical event corpus with meta-knowledge annotation. *BMC Bioinformatics*, 12: 393.

Vincze, V., Szarvas, G., Farkas, R., Mora, G. and Csirik, J. (2008). The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics*, 9(Suppl 11): S9.

Zweigenbaum, P., Demner-Fushman, D., Yu, H. and Cohen, K.B. (2007). Frontiers of biomedical text mining: current progress. *Briefings in Bioinformatics*, 8(5): 358-375.