Enriching Biomedical Events with Meta-knowledge

A thesis submitted to the University of Manchester

for the degree of Doctor of Philosophy

in the Faculty of Engineering and Physical Sciences

March 2013

Raheel Nawaz

School of Computer Science

List of Contents

List of Contents **2**
List of Figures **8**
List of Tables **10**
Abstract **12**
Publications **13**
Declaration **16**
Copyright **17**
Dedication **17**
Acknowledgements **20**
Thesis Outline **21**
Chapter 1: Introduction **23**
1.1 Problem Domain and Motivation 23
1.2 Aims and Objectives 26
1.2.1 Research Aims 26
1.2.2 Hypotheses 27
1.2.3 Research Objectives 27
1.2.4 Research Evaluation 28
1.3 Summary of Contributions 28
Chapter 2: Event-based Biomedical Text Mining **29**
2.1 Introduction to Bio-events 29
2.1.1 Textual Events 29
2.1.2 Bio-events 32
2.1.3 Bio-event Corpora 34
2.2 Bio-event Extraction 37
2.2.1 Shared Tasks 37
2.2.1.1 BioNLP'09 Shared Task on Event Extraction 37
2.2.1.2 BioNLP'11 Shared Task on Event Extraction 39
2.2.1.3 BioNLP'13 Shared Task on Event Extraction 39
2.2.2 State-of-the-Art Systems 40
2.3 Applications of Bio-event Extraction 41
2.3.1 Information Retrieval 42
2.3.2 Linking Pathways to Literature 46
2.3.3 Other Applications 48
2.4 Interpretation of Bio-events 49
Chapter 3: Meta-knowledge **52**

3.1 Need for Meta-knowledge Annotation 52
3.2 Analysis of Related Work 56
3.2.1 Lexical Markers of Meta-Knowledge 56
3.2.2 Existing Corpora with Meta-Knowledge Annotations 59
3.2.2.1 Corpora with Meta-Knowledge Annotations at the Text Span Level 59
3.2.2.2 Corpora with Meta-Knowledge Annotations at the Event Level 61
3.3 Meta-knowledge Annotation Scheme 62
3.3.1 Knowledge Type 63
3.3.1.1 Investigation 64
3.3.1.2 Observation 66
3.3.1.3 Analysis 68
3.3.1.4 Method 72
3.3.1.5 Fact 73
3.3.1.6 Other 74
3.3.2 Certainty Level 76
3.3.2.1 L3 77
3.3.2.2 L2 78
3.3.2.3 L1 79
3.3.3 Polarity 80
3.3.3.1 Positive 80
3.3.3.2 Negative 80
3.3.4 Manner 82
3.3.4.1 High 83
3.3.4.2 Low 84
3.3.4.3 Neutral 85
3.3.5 Knowledge Source 86
3.3.5.1 Current 86
3.3.5.2 Other 88
3.3.6 Hyper-Dimensions 90
3.3.6.1 New Knowledge 90
3.3.6.2 Hypothesis 91
3.4 Hypothetical Annotation Examples 92
Chapter 4: Meta-Knowledge Annotation 97
4.1 Evaluation of the Annotation Scheme 97
4.2 Annotators and Training 97
4.3 Annotation of Abstracts 99
4.3.1 General Corpus Characteristics 99
4.3.1.1 Knowledge Type 100
4.3.1.2 Certainty Level 102
4.3.1.3 Polarity 106

4.3.1.4 Manner....111
4.3.1.5 Knowledge Source....114
4.3.1.6 Hyper-dimensions....115
4.3.2 Inter-Annotator Agreement....116
4.3.3 Annotation Discrepancies....118
4.4 Annotation of Full Papers 122
4.4.1 Knowledge Type 126
4.4.2 Certainty Level 128
4.4.3 Polarity.... 129
4.4.4 Manner.... 130
4.4.5 Knowledge Source.... 131
4.4.6 Hyper-dimensions.... 131
4.5 Comparison of Abstracts and Full Papers 131
4.5.1 Knowledge Type 133
4.5.2 Certainty Level 134
4.5.3 Polarity.... 134
4.5.4 Manner.... 135
4.5.5 Knowledge Source.... 135
4.5.6 Hyper-Dimensions.... 135
4.6 Conclusion 135
Chapter 5: Polarity of Bio-events 137
5.1 Introduction.... 137
5.1.1 Negated Bio-events.... 138
5.1.2 Identification of Negated Bio-events: Task Description and Analysis 139
5.2 Related Work.... 141
5.2.1 Types of Negation.... 142
5.2.2 Negation Cues.... 143
5.2.3 Detection of Negated Terms and Negation Scopes 144
5.2.4 Detection of Negated PPIs.... 145
5.2.5 Detection of Negated Bio-events.... 145
5.3 A Typology of Negated Bio-Events 146
5.3.1 Class Descriptions 147
5.3.1.1 Inherently Negative Bio-events 147
5.3.1.2 Negated Event-trigger.... 148
5.3.1.3 Negated Participant.... 149
5.3.1.4 Negated Attribute.... 150
5.3.1.5 Comparison and Contrast 151
5.3.2 Class Distribution 153
5.3.2.1 Discussion.... 154
5.4 Key Aspects of Negation Detection – Dimensions of Analysis.... 155

5.4.1 Negation Cues 155
5.4.1.1 Ambiguity of Negation Cues 156
5.4.1.2 Indicators of Low Manner of Interaction 157
5.4.1.3 Deactivators of Negation Cues 159
5.4.1.4 Relationship between the Negation Cues and Event-types 160
5.4.1.5 Corpus / Domain Idiosyncrasies 161
5.4.1.6 Compilation of Cue Lists 162
5.4.2 Feature Design 164
5.4.2.1 Semantic Features 166
5.4.2.2 Lexical Features 166
5.4.2.3 Dependency Features 167
5.4.2.4 Command Features 168
5.4.3 Choice of Learning Algorithm 169
5.4.3.1 Decision Trees 169
5.4.3.2 Random Forest 170
5.4.3.3 Logistic Regression 170
5.4.3.4 Naive Bayes 170
5.4.3.5 SVM 171
5.4.3.6 Instance-Based Algorithms 171
5.5 Experimental Settings 171
5.5.1 Datasets 171
5.5.2 Parsing 172
5.5.3 Classifier Implementation 172
5.5.4 Evaluation Measures 173
5.6 Results 173
5.6.1 Best Results for Each Dataset 174
5.6.2 Cue List Comparison 174
5.6.3 Feature Set Comparison 176
5.6.4 Algorithm Comparison 179
5.7 Discussion 181
5.7.1 Comparison with Previous Results 182
5.7.2 Selection of Negation Cues 183
5.7.3 Feature Engineering and Selection 184
5.7.4 Algorithm Selection 186
5.7.5 The Effect of Corpus Size 187
5.7.6 Correlation between Event-Type and Polarity 188
5.8 Conclusion 191
Chapter 6: Manner of Bio-events 193
6.1 Introduction 193
6.1.1 Manner of Bio-Events 194
6.1.2 Annotation of Manner in the Enriched GENIA Event Corpus 197

6.2 Automated Identification of Event Manner 197
6.2.1 Analysis of Manner Cues 197
6.2.1.1 Cue Frequency 198
6.2.1.2 Cue Variation 198
6.2.1.3 Cue Ambiguity 199
6.2.1.4 Combined Event-Triggers / Manner Cues 199
6.2.1.5 Effect of Negation 199
6.2.2 Classifier Design 200
6.2.2.1 Features 200
6.2.2.2 Learning Algorithm 202
6.3 Results and Discussion 202
6.4 Conclusion 204
Chapter 7: Knowledge Source of Bio-events 205
7.1 Introduction 205
7.1.1 Knowledge Source 207
7.1.2 Annotation of Knowledge Source in GENIA-MK and FP-MK Corpora 208
7.2 Analysis of Other Events 209
7.2.1 Cue Frequency 209
7.2.2 Cue Ambiguity 210
7.2.3 Event Complexity 211
7.2.4 Relative Position within Text 212
7.3 Classifier Design 212
7.4 Results and Discussion 213
7.4.1 Abstracts 214
7.4.2 Full Papers 214
7.4.3 Discussion 215
7.5 Conclusion 216
Chapter 8: Meta-knowledge based Discourse Analysis 217
8.1 Introduction 217
8.2 Analysis of Meta-Knowledge Transitions in Abstracts 223
8.2.1 Knowledge Type 224
8.2.1.1 Pair-wise Transitions 224
8.2.1.2 Abstract Level Patterns 230
8.2.2 Certainty Level 233
8.2.2.1 Pair-wise Transitions 233
8.2.2.2 Abstract Level Patterns 236
8.3 Analysis of Meta-Knowledge Transitions in Full Papers 238
8.3.1 Knowledge Type 238
8.3.2 Certainty Level 241
8.4 Conclusion 242

Chapter 9: Conclusion **244**
9.1 Evaluation of Research Objectives and Hypotheses 244
9.1.1 Objective # 1 244
9.1.2 Objective # 2 245
9.1.3 Objective # 3 246
9.1.4 Evaluation of Research Hypotheses 246
9.2 Future Work 247
9.2.1 Meta-knowledge Annotation for Other Domains 247
9.2.2 Meta-knowledge Extraction 248
9.2.3 Discourse Analysis 249

List of Figures

Figure 1. Example of textual event in newswire text 30

Figure 2. Typical representation of the bio-event in sentence S1 33

Figure 3. A simple hypothetical sentence with complex event structure 33

Figure 4. Meta-knowledge Annotation Scheme 63

Figure 5. Inherently negative bio-event – Example 1; Source = PMID: 9427533 147

Figure 6. Inherently Negative Bio-event – Example 2; Source = PMID: 10022882 148

Figure 7. Negated event-trigger – Example 1; Source = PMID: 10022882 149

Figure 8. Negated event-trigger – Example 2; Source = PMID: 790554 149

Figure 9. Negated participant; Source = PMID: 10358173 150

Figure 10. Negated attribute; Source = PMID: 10022882 151

Figure 11. Comparison and contrast – Example 1; Source = PMID: 9427533 152

Figure 13. An instance of the word *loss* with positive contextual (biological) polarity; Source = PMID: 10202937 157

Figure 14. An instance of the low manner indicator little being treated as a negation cue; Source = PMID: 20562282 158

Figure 15. An instance of negation triggered by the construction *no evidence*; Source = PMID: 10221643 160

Figure 16. Cue list comparison: Micro-averaged results for the three datasets 176

Figure 17. Feature set comparison: Micro-averaged results for the three datasets 179

Figure 18. Algorithm comparison: Micro-averaged results for the three datasets 181

Figure 19. Transitions from / to *Knowledge Type* category *Observation* for Abstracts (Abs), Full Papers (FP), and the different sections within full papers, i.e., Background (Back), Results (Res), and Discussion (Disc) 224

Figure 20. Transitions from / to *Knowledge Type* category *Analysis* for Abstracts (Abs), Full Papers (FP), and the different sections within full papers, i.e., Background (Back), Results (Res), and Discussion (Disc) 225

Figure 21. Transitions from / to *Knowledge Type* category *Investigation* for Abstracts (Abs), Full Papers (FP), and the different sections within full papers, i.e., Background (Back), Results (Res), and Discussion (Disc) 225

Figure 22. Transitions from / to *Knowledge Type* category *Fact* for Abstracts (Abs), Full Papers (FP), and the different sections within full papers, i.e., Background (Back), Results (Res), and Discussion (Disc) 226

Figure 23. Transitions from / to *Knowledge Type* category *Method* for Abstracts (Abs), Full Papers (FP), and the different sections within full papers, i.e., Background (Back), Results (Res), and Discussion (Disc) 226

Figure 24. Transitions from / to *Certainty Level* category *L3* for Abstracts (Abs), Full Papers (FP), and the different sections within full papers, i.e., Background (Back), Results (Res), and Discussion (Disc) 234

Figure 25. Transitions from / to *Certainty Level* category *L2* for Abstracts (Abs), Full Papers (FP), and the different sections within full papers, i.e., Background (Back), Results (Res), and Discussion (Disc) 234

Figure 26. Transitions from / to *Certainty Level* category *L1* for Abstracts (Abs), Full Papers (FP), and the different sections within full papers, i.e., Background (Back), Results (Res), and Discussion (Disc) 235

List of Tables

Table 1. Inference Table for New Knowledge Hyper-Dimension **91**

Table 2. Inference Table for Hypothesis Hyper-Dimension **92**

Table 3. Distribution of annotated categories for *Knowledge Type* **100**

Table 4. Most common *Knowledge Type* cue expressions **101**

Table 5. Distribution of annotated categories for *Certainty Level* **102**

Table 6. Most common *Certainty Level* cue expressions **105**

Table 7. Distribution of annotated categories for *Polarity* **106**

Table 8. Distribution of negated events among *Knowledge Type* categories **107**

Table 9. Most common cue expressions for negative polarity **109**

Table 10. Distribution of annotated categories for *Manner* **111**

Table 11. Distribution of negated events among *Knowledge Type* categories **112**

Table 12. Most common *Manner* cue expressions **113**

Table 13. Distribution of annotated categories for Manner **114**

Table 14. Most common cue expressions for *Source=Other* **115**

Table 15. Distribution of categories for the two hyper-dimensions **115**

Table 16. Inter-annotator agreement rates **117**

Table 17. Category distributions for all dimensions **124**

Table 18. Most frequent cues for each category **125**

Table 19. Difference between the category distributions for full papers and abstracts **132**

Table 20. Statistics for bio-event corpora containing polarity information **139**

Table 21. Corpus-wise class distribution of negated bio-events **154**

Table 22. Cue lists **163**

Table 23. Feature sets **165**

Table 24. Best results for each dataset **174**

Table 25. Cue list comparison **175**

Table 26. Feature set comparison **178**

Table 27. Algorithm comparison **180**

Table 28. Classification Results (10-fold CV) **203**

Table 29. Most frequently annotated *Other* cues in GENIA-MK and FP-MK corpora 210

Table 30. Best results for GENIA-MK and FP-MK..... 214

Table 31. Relative frequencies of abstracts starting and ending with events of each *Knowledge Type* category 231

Table 32. Key transition patterns for *Knowledge Type* values in abstracts and their frequencies 231

Table 33. Relative frequencies of abstracts starting and ending with events of each *Certainty Level* category 237

Abstract

Owing to the ever increasing information deluge, it is becoming increasingly difficult to locate relevant information through traditional term-based search methods. Event–based text mining provides a more promising approach, as it also takes into account the semantic relationships between terms.

Typical event representations only focus on identifying the type of the event, its participants and their types. However, additional information, which is essential for correct interpretation of the event, is often present in the text. This includes information about the polarity, certainty level, intensity/rate/frequency, type and source of the knowledge conveyed by the event. We refer to this additional information as *meta-knowledge*.

This thesis focusses on our work involving the enrichment of events with meta-knowledge information. In this thesis we:

- describe the annotation scheme designed specifically to capture meta-knowledge information at the event level
- report on the corpora that have been enriched through deployment of the meta-knowledge annotation scheme
- describe the work on automated identification of meta-knowledge including:
  - a broad-ranging study on analysis and identification of polarity of bio-events using three different bio-event corpora
  - a detailed study on analysis and identification of knowledge source in bio-events found in abstracts as well as in full papers
  - a first study on analysis and identification of bio-event manner
- describe the initial work on a new approach to discourse analysis based on meta-knowledge annotations at the event level

Publications

Intermediate results from this research have been presented and published in the following conferences and journals.

1) **Raheel Nawaz**, Paul Thompson, John McNaught and Sophia Ananiadou (2010). "Meta-Knowledge Annotation of Bio-Events". *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*: Valletta, Malta. pp. 2498-2507.

2) **Raheel Nawaz**, Paul Thompson and Sophia Ananiadou (2010). "Evaluating a Meta-Knowledge Annotation Scheme for Bio-Events". *Workshop on Negation and Speculation in Natural Language Processing (NeSp-NLP 2010), ACL 2010*: Uppsala, Sweden. pp. 69-77

3) Sophia Ananiadou, Paul Thompson and **Raheel Nawaz** (2010). "Improving Search Through Event-based Biomedical Text Mining". *First International Workshop on Automated Motif Discovery in Cultural Heritage and Scientific Communication Texts (AMICUS 2010), CLARIN/DARIAH 2010*: Vienna, Austria.

4) **Raheel Nawaz**, Paul Thompson and Sophia Ananiadou (2010). "Event Interpretation: A Step towards Event-Centred Text Mining". *First International Workshop on Automated Motif Discovery in Cultural Heritage and Scientific Communication Texts (AMICUS 2010), CLARIN/DARIAH 2010*: Vienna, Austria.

5) Paul Thompson, **Raheel Nawaz***, John McNaught and Sophia Ananiadou (2011). "Enriching a biomedical event corpus with meta-knowledge annotation". *BMC Bioinformatics* 2011, 12**:**393

6) **Raheel Nawaz**, Paul Thompson and Sophia Ananiadou (2012). "Identification of Manner in Bio-Events". *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*: Istanbul, Turkey. pp. 3305-3310

7) **Raheel Nawaz**, Paul Thompson and Sophia Ananiadou (2012). "Meta-Knowledge Annotation at the Event Level: Comparison between Abstracts and Full Papers". *Third Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM 2012)*, *LREC 2012*: Istanbul, Turkey. pp. 24-31

8) Maria Liakata, Paul Thompson, Anita de Waard, **Raheel Nawaz** and Sophia Ananiadou (2012). "Comparing Three Models of Scientific Discourse Annotation for Enhanced Knowledge Extraction". *Workshop on Detecting Structure in Scholarly Discourse (DSSD), ACL 2012*: Jeju Island, Korea, pp. 37-46

9) **Raheel Nawaz**, Paul Thompson and Sophia Ananiadou (2013). "Negated Bio-events: Analysis and Identification". *BMC Bioinformatics 2013*, 14:14

10) **Raheel Nawaz**, Paul Thompson, Sophia Ananiadou (2013), "Towards Event-Based Discourse Analysis of Biomedical Text". *CICLing 2013*: Samos, Greece.

11) **Raheel Nawaz**, Paul Thompson, Sophia Ananiadou (2013), "Something Old, Something New: Identifying Knowledge Source in Bio-Events". *CICLing 2013*: Samos, Greece.

12) Riza Batista-Navarro, Georgios Kontonatsios, Claudiu Mihăilă, Paul Thompson, **Raheel Nawaz**, Ioannis Korkontzelos, Sophia Ananiadou (2013), "Supporting Discourse Phenomena in an Interoperable NLP Framework". *CICLing 2013*: Samos, Greece.

* -- *joint first author*

Other work carried out as indirect part of this research has been presented and published in the following conferences and journals.

1) Xinglong Wang, Rafal Rak, Angelo Restificar, Chikashi Nobata, C.J. Rupp, Riza Theresa B. Batista-Navarro, **Raheel Nawaz** and Sophia Ananiadou (2010). "NaCTeM Systems for BioCreative III PPI Tasks". *BioCreative III Workshop (BioCreative 2010)*: Bethesda, MD, USA.

2) Xinglong Wang, Rafal Rak, Angelo Restificar, Chikashi Nobata, C.J. Rupp, Riza Theresa B. Batista-Navarro, **Raheel Nawaz** and Sophia Ananiadou (2011). "Detecting experimental techniques and selecting relevant documents for protein-protein interactions from biomedical literature". *BMC Bioinformatics* 2011, 12:8

Declaration

No portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright

ter.ac.uk/library/aboutus/regulations) and in The University's policy on presentation of Theses.

Dedication

To Zuna and Daniyal

Acknowledgements

Arguably, the most important part of a successful PhD project is the student-supervisor relationship. While it is *good* to have a supervisor who is famous and prestigious, it is *great* to have one who is also kind and caring. This is where I feel that I have been tremendously lucky to have **Professor Sophia Ananiadou** as my supervisor. In the past four years, Sophia has simultaneously acted as my teacher, mentor, critic, coach and advisor. She encouraged me every step of the way, and was especially considerate when I hit a rough patch. I cannot thank her enough for her guidance, support and kindness.

I have constantly felt that the National Centre for Text Mining (NaCTeM) provides an ideal environment for research. While I have always found all team members to be approachable and helpful, I would especially like to thank **Mr Paul Thompson**. Paul has been a superb research partner, whose friendship is as valuable as his research input.

My quest for a PhD in computer science started nearly seven year back, when I enrolled (part-time) for the MSc in Advanced Computer Science. This has been a long and arduous journey that demanded tough balancing between the family, study and work commitments. None of this would have been possible without **Zuna**, my soul mate, who made countless adjustments and sacrifices to allow me to peruse my dream.

I must also thank **my parents** and **parents-in-law** for their continuous encouragement and motivation.

Finally, I want to thank **Daniyal**, my son, whose smile is the best antidote to stress.

Thesis Outline

This thesis contains the following chapters:

- **Chapter 1** provides an introduction to the research project and describes the problem domain, motivation, aims and objectives of the research.
- **Chapter 2** provides a brief introduction to event-based biomedical text mining. It starts with a description of bio-events, followed by discussions on bio-event extraction systems and the key applications of such systems. The chapter concludes with a discussion on the limitations of current bio-event representations in terms of providing complete semantic interpretations.
- **Chapter 3** discusses the meta-knowledge annotation scheme in detail. It starts with a discussion on the need for meta-knowledge annotation at the level of bio-events. This is followed by a detailed description of the annotation scheme with several examples.
- **Chapter 4** provides an overview of the application of the meta-knowledge annotation scheme to two bio-event corpora. The annotation results for both corpora are discussed in detail and a comparison is provided.
- **Chapter 5** describes the work on first detailed study on the analysis and identification of negated bio-events. The analysis is based on the negated events in three open access bio-event corpora. The chapter begins with a detailed introduction to the task. This is followed by analyses of the types of negated bio-events and the three main aspects of the problem of automated identification of negated bio-events, i.e., negation cues, feature design and

choice of learning algorithm. Finally, the experiments and results are discussed.

- **Chapter 6** describes the work on the analysis and identification of bio-event manner. The chapter begins with an introduction to manner annotation for bio-events. This is followed by an analysis of manner cues, and a discussion on classifier design. Finally, the results of classification experiments are discussed.

- **Chapter 7** describes the work on the analysis and identification of knowledge source for bio-events. It starts with a detailed discussion on annotation of knowledge source at different levels of granularity. This is followed by the analysis of events attributed to explicit external sources, and a discussion on classifier design. Finally, the results of classification experiments are discussed.

- **Chapter 8** provides an overview of the initial work on discourse analysis based on meta-knowledge annotation at the event level. It focusses on two dimensions of meta-knowledge, i.e., *Knowledge Type* and *Certainty Level*. The analysis includes both local and global transition patterns observed in abstracts and full papers.

- **Chapter 9** provides an evaluation of the research work against the aims and objectives described in chapter 1. It concludes with a discussion on future work

Chapter 1: Introduction

This chapter provides an introduction to the research project. It begins with an introduction to the problem domain and the motivation for this research. This is followed by an outline of research aims and objectives.

1.1 Problem Domain and Motivation

Biomedical research literature is being published at an ever-increasing rate. MEDLINE[1], which is the largest repository of biomedical research literature, already contains over 19 million citations and between 2,000 to 4,000 new entries are being added to it every day. This makes it highly important to provide researchers with automated, efficient and accurate means of locating the information they require, allowing them to keep abreast of developments within biomedicine [1-6]. However, searching using keywords alone will usually return far more documents than are relevant to a query. For example, a researcher interested in finding which proteins are *positively regulated* by *IL-2*, would typically expect the following sentence answering his/her query:

S1 *These results suggest that p21ras proteins are activated by IL-2 in normal human T lymphocytes.*

Using the search terms *IL-2* and *activate* on a typical search engine would return a long list of results. Although some documents containing information directly relevant to the user's query (e.g., the document containing S1) would be amongst the retrieved results, it is highly likely that many of the retrieved documents will not contain the required information. This is because a simple keyword query cannot

[1] http://www.nlm.nih.gov/databases/databases_medline.html

express the fact that the user requires there to be a particular semantic relationship between the two search terms, i.e. only documents in which *IL-2* is expressed as the instigator of the *positive regulation* are of interest. For the verb *activate*, the instigator corresponds to the grammatical subject. Generally, search engines view documents as "bags of words" that have no internal structure. Thus, there is no guarantee that in the returned documents, there will be any kind of relation between the search terms. Documents will still be returned even if the terms are entirely unrelated, but nonetheless exist somewhere in the text.

Since most biomedical terms have variant forms, the above query is also likely to fail to return some relevant documents. For example, *IL-2* is the short form for *interleukin-2*, and both forms may appear in text either with or without the hyphen. Moreover, variants are not restricted to acronyms/full forms and minor orthographic variations, but may include synonymous terms that are completely unrelated. For example, *T-cell growth factor* is often used as a synonym of *interleukin-2*, and this term also has its own variant forms (e.g., *TCGF*). A further issue relating to the above query is that the user is interested specifically in retrieving information about biological reactions that correspond to *positive regulations*. The verb *activate* is a common means of describing such reactions in text. However, this is by no means the only way in which *positive regulations* can be described; there are also many possible variations in this respect, such as the use of other verbs, e.g., *stimulate* or *affect*, or using nouns that convey a similar meaning, e.g., *activation, activator, effect, stimulation*, etc. A typical search engine would not find all such variants automatically. Although a query could be formulated to include some variants, this is cumbersome and time-consuming for the user, and it would be extremely difficult to enumerate all possible variants.

The above limitations of search engines can be alleviated through the integration of text mining methods [1, 5, 7]. In particular, the use of event extraction systems can facilitate the development of event-based search systems. Events are structured, semantic representations of pieces of knowledge contained within a text. In the biomedical field, they may include various biological processes, such as regulation, expression and transcription, whilst examples from newswire include terrorist attacks, company takeovers, personnel appointments, etc. In event-based search systems, searches take place over these structured events, rather than over unstructured text.

Although event-based searching can retrieve many more relevant documents than is possible using traditional keyword searches, the typical event representations (and the event extraction systems based on such representations) do not take into account all available information pertaining to the interpretation of the event. For example, a particular event may represent generally accepted knowledge, experimental observations, hypotheses or analyses of experimental results. For the two latter types of event, the author may express varying degrees of certainty regarding the analysis performed. Similarly, other interpretative information about the event is also often available in the text. This includes: the information about the polarity of the event (i.e., whether the event is negated or not), the manner in which it takes place (i.e., the intensity, rate and frequency of the event), and the knowledge source to which the event can be attributed. We term these types of interpretative information collectively as *meta-knowledge* [8].

Without access to meta-knowledge, a large number of extracted bio-events will be treated identically by text mining systems, even though their intended interpretations may vary significantly [8, 9]. This poses a serious problem to users of the system

whose information requirements include the ability to distinguish between certain interpretations. For example, a biologist who wishes to update either an incomplete model of a biological process (e.g., a molecular pathway [10]) or a curated biological database [11] would wish to locate only newly-reported, reliable experimental knowledge. Thus, he would be interested only in experimental observations or confident analyses of results, but not in hypotheses or more tentative analyses. Similarly, certain users may be interested specifically in negated interactions, whilst others may want to exclude them from their retrieved results. Further cases where interpretation can be important include matching hypotheses with experimental observations/evidence, or detecting contradictions that occur in the literature [12].

In this thesis, we report on the research carried out to evaluate the feasibility of identifying meta-knowledge information at the event level. We firstly present our annotation scheme for enriching bio-events with meta-knowledge information, and report on the application of the scheme to existing corpora containing event annotations. Subsequently, we describe our efforts to train systems to recognise meta-knowledge information at the event level automatically. Finally, we report on a new approach to discourse analysis based on meta-knowledge annotations at the event level.

1.2 Aims and Objectives

A brief description of our project research aims, objectives, hypotheses and evaluation measures is as follows:

1.2.1 Research Aims

We refer to the overall aim of our research as A_O, and define it as follows:

A_O	*To investigate the feasibility of an event-centred approach for meta-knowledge annotation and extraction*

We refer to the specific aims of our research as A_{Sn}, and define them as follows:

A_{S1}	*To identify the necessary information required for correct interpretation of bio-events*
A_{S2}	*To develop a methodology for recording the information required for correct interpretation of bio-events*
A_{S3}	*To develop a methodology for extracting the information required for correct interpretation of bio-events*

1.2.2 Hypotheses

The research effort is being driven by the following main hypotheses:

H_1	*Discrete information about event interpretation can be identified – Meta-knowledge annotation can be performed at the event level*
H_2	*The above information can be automatically extracted – Meta-knowledge can be extracted automatically*

1.2.3 Research Objectives

The research objectives are as follows:

O_1	*To develop an annotation scheme for capturing the information necessary for the correct interpretation of bio-events*
O_2	*To develop manually annotated corpora of bio-events with the required interpretative information*

O_3	*To develop automated systems for enriching bio-events with the required interpretative information*

1.2.4 Research Evaluation

Three traditional evaluation methodologies will be followed for assessing the quality of resources produced in the course of this research:

E1	*The annotation schemes and their corresponding corpora will be evaluated using inter-annotator consistency and agreement rates.*
E2	*The automated systems will be evaluated using the traditional measures of precision, recall and F-mesaure.*
E3	*The methodologies will be evaluated based on the performance of their corresponding systems.*

1.3 Summary of Contributions

The research presented in this thesis has made the following contributions:

- Development of a meta-knowledge annotation scheme and annotation guidelines [8, 13-17]
- Creation of two meta-knowledge enriched corpora of bio-events: GENIA-MK [9] and FP-MK [18]
- The first comprehensive study on the analysis and identification of negated bio-events [19]
- Development of a system for automated identification of event manner [20]
- Development of a system for automated identification of event knowledge source [21]
- Initial work on meta-knowledge based discourse analysis [22]

Chapter 2: Event-based Biomedical Text Mining

This chapter provides an introduction to event-based biomedical text mining. It starts with an introduction to bio-events, followed by discussions on bio-event extraction systems and the key applications of such systems. The chapter concludes with a discussion on the limitations of current bio-event representations in terms of providing complete semantic interpretations.

2.1 Introduction to Bio-events

2.1.1 Textual Events

In its most general form, a textual event can be described as an action, relation, process or state expressed in the text [23]. More specifically, a textual event is a structured semantic representation of a certain piece of information contained within the text. Textual events are usually anchored to particular text fragments that are central to the description of the event. The most important of these text fragments is the *event-trigger*, which is usually a verb or a noun that indicates the occurrence of the event. Events are often represented by a template-like structure with slots that are filled by the event *participants*. These event participants describe the different aspects of the event, e.g., what caused the event, what is affected by it, where it took place, etc. Based on its function, each participant can be assigned a semantic role within the event. The participants can correspond to entities, concepts or even other events. If an event contains one or more events amongst its participants, then it is called a *complex* event. Typically each event is also assigned a type/class from an

event taxonomy/ontology. Similarly, the entities participating in the event are also assigned types/classes from an entity taxonomy/ontology.

As an example, consider Figure 1. It shows a sentence from general newswire text which contains an event relating to the founding of an organisation. The figure also shows the event representation (as per the ACE guidelines [24]). The event is centred on the word *founded* (which has been identified as the *event-trigger*), and it has been an event type of *Start_Org*. Three participants have been identified for the event: *Joseph Conrad Parkhurst* (who has been assigned an entity type of *Person*) plays the role of *Agent*; *Cycle World* (which has been assigned an entity type of *Organisation*) plays the role of *Org*; and the year *1962* has been identified as the *Time* of the event.

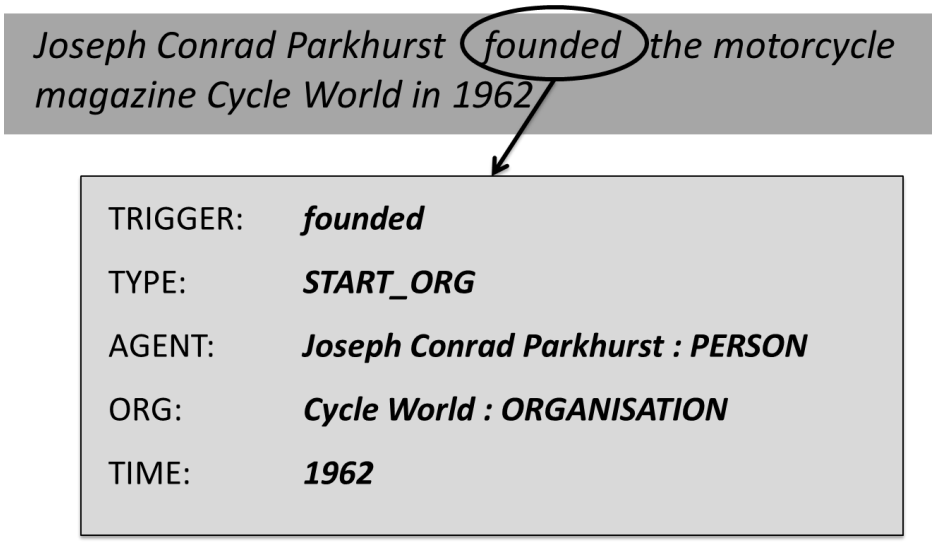

Figure 1. Example of textual event in newswire text

The event representation of text allows a document to be viewed as a collection of nested events. We call this the **event view** of a document. This event-view has some similarity with the document view based on *atomic propositions* as defined by Akhmatova [25]:

> *An atomic proposition is a minimal declarative statement (or a small idea) that is either true (T) or false (F) and whose truth or falsity does not depend on the truth or falsity of any other proposition.*

Depending upon the granularity of events, the event-view can capture most of the atomic propositions. The event-view also has some similarity with the document view based on *discourse commitments* proposed by Hickl & Bensley [26]:

> *Discourse commitments represent the set of propositions which can necessarily be inferred to be true given a conventional reading of the text*

The discourse-commitment-view ascribes to a given text fragment even those propositions that are not explicitly mentioned, but which can be inferred from the text fragment, e.g., conventional implicatures and conversational implicatures, etc. The event-view, in contrast, can usually only capture the explicitly mentioned propositions. However, when used in conjunction with event and term ontologies, the event-view can facilitate the extraction of inferred discourse commitments.

Finally, it is important to note that although the general format of event templates can be comparable across different domains, the features of the events to be recognised vary in several ways, i.e., both in terms of the event types to be extracted and the types/roles of participants to be recognised. Therefore, different event representations are required for different domains. Furthermore, many subdomains can exist within a specific domain, and each of these subdomains can have its own sublanguage and informational structure [27]. Therefore, a generic analysis cannot capture the informational structure of a sublanguage [28], which demands richer relations expressing conditions, manner, destination, etc. Sublanguage-driven information extraction systems rely on the notion that the informational structure of the

domain imposes constraints at all linguistic levels (lexical, syntactic, semantic, and discourse), which can be exploited to produce accurate systems. Therefore, different event representations can be required for different subdomains within a domain.

2.1.2 Bio-events

A bio-event is a textual event specialised for the biomedical domain. Kim et al [29] define a bio-event as "a dynamic bio-relation involving one or more participants". These participants can be bio-entities or (other) bio-events, and are each assigned a semantic role like *theme* and *cause*, etc. Each bio-event is typically assigned a type/class from a chosen bio-event taxonomy/ontology, e.g., the GENIA Event Ontology [29]. Similarly, the bio-entities are also assigned types/classes from a chosen taxonomy/ontology, e.g., the Gene Ontology [30]. The template of a bio-event can also contain additional slots, e.g., to denote temporal and spatial attributes.

As an example, consider the sentence S1:

S1 *These results suggest that p21ras proteins are activated by IL-2 in normal human T lymphocytes.*

This sentence contains a single bio-event, anchored to the verb *activates*. Figure 2 shows a typical structured representation of this bio-event. The fact that the event is anchored to the word *activates* allows the event-type of *Positive Regulation* to be assigned. The event has two slots, i.e. *theme* and *cause* whose labels help to characterise the contribution that the slot filler makes towards the meaning of the event. In this case, the slots are filled by the subject and object of the verb *activate*, both of which correspond to the same type of bio-entities (i.e., *Protein)*.

TRIGGER:	***activates***
TYPE:	**POSITIVE REGULATION**
THEME:	***p21ras proteins :*** **PROTEIN**
CAUSE:	***IL-2*** **: PROTEIN**
LOCATION:	***normal human T lymphocytes*** **: CELL**

Figure 2. Typical representation of the bio-event in sentence S1

Figure 3 shows a simple hypothetical sentence with a more complex event structure. The event E1 is anchored to the word *expression* and has been assigned the event type of *Gene Expression*. It has a single participant, the arbitrary gene *X*, which acts as the theme of the event. E1 also has a location attribute, which has the arbitrary value of *Z*. The word *activates* has been identified as the event-trigger for the complex event E2, which has been classed as a *Positive Regulation* event. It has two participants: the arbitrary protein *Y* and the event E1, which act as the cause and the theme of the event, respectively.

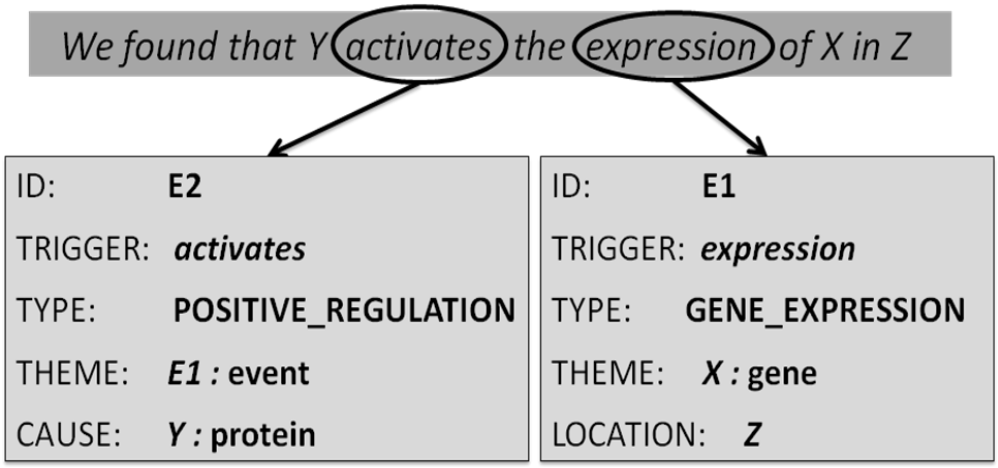

Figure 3. A simple hypothetical sentence with complex event structure

Relationship between Bio-events and Other Types of Bio-relations

The above definition of bio-event has been used as the basis for various annotation and extraction tasks [29, 31-34]. However, it is important to note that this is a fairly general definition; i.e., it is much broader in scope than the other types of bio-relations which have received significant attention in the past. The most notable of these bio-relations is Protein-Protein Interaction (PPI), which is generally defined as "an instance of the mention of an interaction between two proteins" [35, 36]. The structured representation of a PPI is simpler and coarser than that of a bio-event. More interestingly, a PPI can be viewed as a special case of bio-event, where the participants are restricted to proteins. Therefore, every instance of a PPI is a bio-event, while the converse is not true. Similarly, other (more specific) types of bio-relations can also be viewed as special cases of bio-events, for example, genotype-phenotype associations [37, 38], disease-gene associations [39, 40], etc.

Historically**,** different bioinformatics tasks have motivated the extraction of different types of bio-relations. For example, PPI extraction has been motivated by the need to populate interaction databases, such as MINT [41]. However, bio-event extraction aims to support the development of richer, more detailed and more structured databases, like Pathguide [42] and Gene Ontology Annotation [43]. A detailed discussion of this topic can be found in [29, 44].

2.1.3 Bio-event Corpora

Annotated bio-event corpora are a vital resource for the development of event-based text mining systems. These corpora provide direct evidence of how events manifest themselves in texts, and as such, they can be used in both the development and train-

ing of event extraction systems, as well as in the evaluation of the performance of such systems, by acting as a "gold standard" [45].

Recently, significant effort has been put into the creation of various bio-event corpora. Although each one of these corpora has been created with different aims and motivations, they all contain bio-events of varying levels of granularity [32, 44]. A brief description of some of these bio-event corpora is given below.

GENIA Event

The GENIA Event corpus [29] contains 1,000 MEDLINE abstracts in which 36,858 bio-events have been identified. Each event belongs to one of the 36 event classes defined in the GENIA Event Ontology [29]. The event participants can be bio-entities or other bio-events. Each bio-entity belongs to one of the 46 classes defined in the GENIA Term Ontology [29]. Other than the participants, an event may contain additional attributes including location, time and experimental context.

BioInfer

The BioInfer [31] corpus contains 1,100 sentences in which 2,662 bio-events have been identified. Each event belongs to one of the 60 event classes defined in the BioInfer Relationship Ontology [31]. It is important to note that a more general definition of bio-event has been used in BioInfer, and static bio-relations [46] have also been marked as bio-events.

BioNLP'09 ST

The BioNLP'09 ST corpus [47] is a modified subset of the GENIA Event corpus, which was created for the BioNLP'09 shared task on event extraction (further details in section 2.2.1). It contains 950 MEDLINE abstracts, which are divided into two

subsets: the *Development* subset, comprising 150 abstracts and the *Training* subset comprising 800 abstracts. The corpus contains a total of 11,480 bio-events, and each bio-event belongs to one of 9 event classes from the GENIA Event Ontology.

BioNLP'11 ST

The BioNLP'11 ST corpus [48] is an extended version of the BioNLP'09 ST corpus, which was created for the BioNLP'11 shared task on event extraction (further details in section 2.2.1). It contains the entire BioNLP'09 ST corpus with various additional entity and event type annotations. It also contains a new subset of five full papers annotated with 3,150 bio-events.

GREC

The Gene Regulation Event Corpus (GREC) [49] contains 240 MEDLINE abstracts in which 3,067 bio-events have been identified. Each event has a set of arguments, which can include both the event participants and attributes like time, location and manner etc. The bio-events and their participating bio-entities have been assigned classes from the Gene Regulation Ontology [50].

GeneReg

The GeneReg [51] corpus contains 314 MEDLINE abstracts in which 1,770 bio-events have been identified. Each event belongs to one of the 4 classes from the Gene Regulation Ontology [50].

2.2 Bio-event Extraction

2.2.1 Shared Tasks

Shared tasks bring together different research teams to focus on a problem by providing standard datasets and a common evaluation framework. They focus the attention of the research community on timely issues and act as a driver for the specification of new tasks and challenges [44]. Within the domain of biomedical text mining, shared tasks have played a significant role in advancing the state-of-the-art in various types of systems [47, 52]. For example, the TREC Genomics track [53] focussed on information retrieval whilst the Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA) [54] targeted named entity recognition. In terms of relation extraction, the Learning Language in Logic (LLL) challenge [55] and the BioCreative challenges [52] have yielded significant progress. However, both LLL and BioCreative have focussed on simple representations of relations between bio-entities, i.e., protein-protein interactions. A step towards recognition of more detailed and intricate representations of bio-relations (i.e., bio-events) has been the introduction of BioNLP shared tasks on event extraction [33].

2.2.1.1 BioNLP'09 Shared Task on Event Extraction

The first of the BioNLP shared tasks, i.e. BioNLP'09 [33], was based on a dataset that was largely derived from the GENIA Event corpus. However, since this was the first shared task of its type, the bio-event data was simplified to ensure that the task remained tractable. Specifically, the BioNLP'09 ST corpus (section 2.1.3) contains around a third of the events in the GENIA Event corpus, and only 9 event types

(compared to 36 types in the GENIA Event corpus). The core task involved locating bio-event triggers, assigning event types, and identifying the main event participants. The optional tasks included identification of additional information about the bio-event including temporal and spatial information, and the negation and speculation status of the event.

A total of 42 teams showed interest in the shared task and registered for participation, and 24 teams submitted final results. All 24 teams participated in the core task, while, six teams participated in each of the optional tasks. The complexity of the task was indicated by the composition of participating teams, which included computer scientists, bioinformaticians, biologists and linguists.

The evaluation of submitted systems showed a broad performance range. However, the results were both promising and encouraging for the future of bio-event extraction [47]. For simpler events (i.e., events with only one primary participant, e.g., *Gene Expression*, *Transcription*, *Phosphorylation*, etc.), the top-ranked systems achieved F-scores of around 70%. This was particularly encouraging, as systems with such performance levels could be further enhanced for practical applications. However, the evaluation results for more complex events (i.e., events with more than one primary participant, e.g., *Binding*, *Regulation* and its subcategories: *Positive Regulation* and *Negative Regulation*, etc.) were significantly lower with the top-ranking systems achieving F-scores of 40-45%. This showed that the extraction of such events is much more challenging and requires further analysis.

2.2.1.2 BioNLP'11 Shared Task on Event Extraction

In the second BioNLP shared task, i.e., BioNLP'11 [56-58], the number of tasks (and with it, the range of event annotated corpora made available) increased considerably. Building on the success of the BioNLP'09 shared task, the main objective of the BioNLP'11 shared task was to extend the scope of bio-event extraction by diversifying the text types, subject domains and event types under consideration. In terms of text types, the previous efforts at event extraction in the biomedical domain had been almost exclusively restricted to abstracts. The BioNLP'11 shared task introduced the recognition of events in full papers. This is considered vitally important for scalable event extraction systems, given that, on average, less than 8% of the scientific claims of a complete paper occur in the abstract [59]. Regarding subject domains, the BioNLP'09 shared task focussed on a single subdomain of molecular biology (i.e., human transcription factor in blood cells). The BioNLP'11 shared task included event extraction from three additional subdomains (i.e., two-component systems, bacteria biology and bacillus subtilis). Similarly, the BioNLP'09 shared task only considered nine types of bio-events. The BioNLP'11 shared task includes five different tasks of event extraction with a total of 46 event types. A further focus of this shared task was to evaluate the performance of systems on their ability to carry out supporting tasks that are considered essential to allow advances in the performance of event extraction systems, for example, resolution of co-reference between entities.

BioNLP'11 shared task received a total of 46 submissions from 24 teams. In comparison to the BioNLP'09 shared task, a 10% overall reduction in error rate was observed, with a significant improvement in the ability of systems to recognise more

complex bio-events. The evaluation results also showed that generalisation to full papers is feasible, with a modest loss in performance compared to abstracts. Similarly, it was noticed that the removal of subdomain specificity does not compromise extraction performance [34, 48].

2.2.1.3 BioNLP'13 Shared Task on Event Extraction

The third BioNLP shared task, i.e., BioNLP'13[2], has recently been announced. While this shared task follows the general outline and goals of the previous tasks, it broadens the scope of biomedical text-mining applications by introducing new tasks on cancer genetics and pathway curation. Moreover, it takes a step further to include construction of knowledge bases by linking event extraction with semantic web, ontology population, and pathway construction technologies.

2.2.2 State-of-the-Art Systems

The successive BioNLP shared tasks on event extraction have resulted in the development and improvement of various bio-event extraction systems. This has significantly improved the state-of-the-art in this area. Brief descriptions of some of these systems are as follows:

EventMine

EventMine [60] is a state-of the-art event extraction system. It is similar to the types of systems that have appeared since the initiation of the BioNLP shared tasks on event extraction, i.e., it can extract semantically-oriented events that conform to the bio-event template introduced above (section 2.1). EventMine has been shown to be

[2] http:// 2013.bionlp-st.org/

particularly strong in identifying events with extended sets of arguments, since it is able to outperform all systems that participated in the BioNLP'09 ST. Recently, the system has been further refined by employing domain adaptation and coreference resolution [61].

Turku Event Extraction System (TEES)

The Turku Event Extraction System (TEES) [62] is a versatile and scalable event extraction system. It achieved the best overall performance in the BioNLP'09 shared task on event extraction. An updated version of the system [63] was submitted for the BioNLP'11 shared task on event extraction. The generalisability of the system, in terms of its ability to extract bio-events from different subdomains with different event types, was demonstrated by its successful application to all tasks and subtasks (with top performance in several tasks) within the shared task. Recently, TEES has also been deployed to extract events from 18 million PubMed abstracts [64].

FAUST

FAUST [65] is a state-of-the-art event extraction system, which achieved the best overall performance in three tasks in the BioNLP'11 shared task on event extraction. Compared to the other event extraction systems, FAUST has a unique architecture in the sense that instead of using a single (machine learning) model to produce its output, it uses a combination of multiple models, where the output of one model is used as additional features in another model.

2.3 Applications of Bio-event Extraction

Automatic extraction of bio-events has a broad range of applications [44], from support for the creation and annotation of pathways [10, 66] to automatic population or

enrichment of databases [11]. However, the most apparent and broad ranging application of bio-event extraction is in semantic information retrieval systems.

2.3.1 Information Retrieval

Text mining systems that are able to extract events automatically can allow much more precise and focused retrieval and extraction than the traditional keyword-based systems [17]. Event-based retrieval allows the user to specify one or more constraints on the events to be retrieved, which are not dependent on the precise wording in the text. These constraints could be in terms of the type of the event, and/or the type of its participants, and/or the value of a participant in a particular role. An example of such a system is MEDIE [67], which is a semantic search engine that facilitates structured, event-based searching over MEDLINE abstracts. It is currently configured for searching biomedical documents. However, the general architecture could be adapted to other domains through substitution/adaptation of the various modules. The modules include a deep syntactic analyser that is tuned to the biomedical domain [68], an event expression recogniser and a named entity recogniser [69]. Queries take the form of <*subject, verb, object*> to specify an event, where *subject* and *object* refer to grammatical relations with the verb. Such relations often hold between the primary participants of events. For example, in the biomedical event example in sentence S1 (section 2.1), the subject (i.e., *IL-2*) corresponds to the *Cause*, whilst the object (i.e., *p21ras proteins*) corresponds to the *Theme.* One or more of the three "slots" in the query template can be left empty, in order to increase or decrease the specificity of the query. For example, the query to find out which proteins are positively regulated by *IL-2* would be encoded as follows: <*IL-2, activate, ?*>.

MEDIE addresses the issues of the simple keyword-based searching, at least to a certain extent:

- Only documents in which the specified grammatical relations hold between the search terms are retrieved, thus eliminating many of the spurious results retrieved by a traditional search engine. In contrast, the use of deep parsing technology [70], allows MEDIE to retrieve even those documents that contain grammatical variants of the specified query, e.g., active or passive voice constructions.

- MEDIE detects named entities and event-trigger terms, which are then linked with databases and ontologies. This allows the automatic expansion of searches to include variants of search terms that are listed in these resources. Named entities in the *subject* and *object* slots of the event template are linked with the Unified Medical Language System (UMLS) meta-thesaurus[3], whilst variants of verbs are retrieved via linking with the Gene Ontology [30].

- Each sentence is automatically classified by MEDIE as *title, objective, method, result* or *conclusion*, and searches can specify which of these sentence types to consider when retrieving results. For example, events in *result* sentences are likely to contain definite experimental results, whilst *conclusion* sentences will usually contain analyses or conclusions about experimental results. This allows some level of control over restricting the discourse or meta-knowledge contexts in which retrieved events occur.

[3] http://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/index.html

Despite its clear advantages over a traditional search engine, MEDIE still presents some limitations. Firstly, it only allows the specification of two event participants, i.e., the subject and object of the verb. Although these usually constitute the major participants of the event, there are frequently other participants, e.g., location and time. In biomedical texts in particular, information corresponding to time, environmental conditions and manner is considered to be highly important to their correct interpretation [71]. It is thus useful to allow users to specify restrictions on a greater range of event participants.

A further potential issue with MEDIE is that its search template is closely tied to the syntactic structure of the text. However, for several reasons, a search approach in which users specify restrictions in terms of semantic rather than grammatical roles is more desirable. For instance, the *Cause* and *Theme* semantic arguments do not consistently correspond to the grammatical subject and object for all verbs. A semantic approach is even more desirable if additional participants (e.g., location, environmental conditions, etc.) are taken into account and may be specified as part of the search criteria. Several of these participant types are specified through syntactically similar means, i.e., through the use of prepositional or adverbial phrases [72].

A further restriction of MEDIE is that it can only retrieve events whose triggers are verbs. Given the prevalence of nominalised forms in biomedical texts [73], e.g., *activation* rather than *activate*, many relevant events may be missed.

MEDIE's search strategy is largely based on syntactic analysis of text. Whilst this is a vast improvement over the usual bag-of-words approach, events are semantic rather than syntactic structures. Although analysing syntactic structure is a prerequisite for extracting semantic event structures, this step should preferably be kept "behind

the scenes" in event-based search systems. By allowing specification of search criteria through an intuitive semantic template that is independent of the exact textual manifestation of events, users without linguistic expertise can be empowered to perform sophisticated semantic searches. An ideal template would allow the specification of the following types of search options:

- Specification of event types (chosen from a fixed set) as an alternative to specifying specific event-trigger words or phrases. Use of hierarchically-structured ontologies of event types can provide the user with control over the level of generality of the results returned by the query.
- Use of semantic role types rather than grammatical relations when specifying restrictions on event participants.
- A flexible way of specifying restrictions on the values of particular participants, in the form of either terms (e.g. *NF-kappa B*), term classes (e.g. *Protein*), or a combination of both. Again, hierarchically-structured sets of terms can give the user control over the specificity of the results returned by the query.
- Specification of interpretation information about the event, e.g., should only facts be retrieved or are experimental analyses also acceptable? If so, are highly speculative analyses of interest, or only more definite analyses?

The main challenges of producing a system that can extract events that will match such a template are as follows:

1) Identification of event-triggers and their respective ontological types
2) Identification of event participants (terms or other events), their semantic roles and their respective ontological types

3) Identification of the contextual information required for the correct interpretation of the event

State-of-the-art event extraction systems (section 2.2) address the first two challenges by using a combination of sophisticated parsing technologies and event/entity ontologies. For example, EventMine uses the Enju parser [70], the GDep parser [74], the GENIA event ontology [29], and the GENIA term ontology [29]. However, the third challenge, i.e., the identification of contextual information necessary for the correct interpretation of the event, remains an understudied area. This issue is further discussed in sections 2.4 and 3.1.

2.3.2 Linking Pathways to Literature

Biochemical signalling and metabolomic pathways are becoming increasingly important for biomedical research because they represent collective interpretations of facts scattered throughout literature [75-78]. Owing to the very integrated nature of pathways, they require substantial human effort to construct, i.e., researchers have to read a large number of published papers, interpret them and construct a pathway [79]. The curation of a constructed pathway also requires monitoring of recent publications in order to maintain relevance. Furthermore, since different interpretations of the same set of facts are possible, researchers often want to read the original papers from which a pathway is constructed, to ensure it is carried out in a manner consistent with their interpretation [76, 80]. Therefore, researchers can benefit considerably from the use of text mining tools, not only to support the maintenance of pathway models [81], but also to provide direct links from these pathways to the supporting evidence in relevant literature [82]. Furthermore, such tools can also help

in keeping existing pathway models up to date by revising them according to newly published articles.

Linking pathways to literature evidence and aiding pathway construction and enrichment are two of the most important applications of event recognition to systems biology [10]. Some previous studie involving text mining technology for pathway construction have focused on extracting binary interactions between proteins or genes [83-85]. Although the resultant networks seem to be pathways, they do not represent any coherent interpretations of the reported facts [10]. Mapping between the results of automatically constructed networks and pathways requires a deeper analysis that emulates the interpretations of biologists, including inferences based on biological background knowledge. Thus, providing evidence from the literature about pathway representations requires the extraction not only of events, but also of the relevant context around them [44].

PathText [66] is an event-based integrated environment for biological pathway visualisation, which brings together the strengths of different text mining tools, including advanced searches based on event extraction. PathText links several text mining systems (including FACTA [86], KLEIO [87] and MEDIE [67]) to provide a flexible interactive environment which allows a researcher to navigate from pathway visualisation to text mining, to retrieve recently published articles which are potentially relevant, to browse them and to associate them with relevant parts of pathways.

Although the incorporation of event extraction makes PathText one of the most sophisticated and versatile pathway construction tools available [44], the underlying event extraction technology suffers from the same limitations as those of MEDIE (section 2.3.1). Most particularly, the event extraction technology used in PathText

does not consider the necessary contextual information required for the correct interpretation of bio-event. The task of pathway construction can be further helped by enhancing the underlying event extraction technology to automatically identify this contextual information. This would enable the system to automatically identify the subtle epistemic aspects of an event; for example, its polarity, certainty level, type and source of knowledge being conveyed.

2.3.3 Other Applications

Finding Implicit Associations between Entities

FACTA [86] is an interactive text-mining system designed to help researchers find both explicit and implicit associations between biomedical concepts over the entire MEDLINE corpus. It is capable of producing ranked lists of important biomedical concepts, e.g. genes, diseases and chemical compounds, which are considered relevant to the query according to their co-occurrence statistics. FACTA + [88] is an enhanced version of the original system, which incorporates additional features including the use of bio-event extraction in the underlying search system. This allows the user to search for implicit/explicit associations in documents containing specified bio-entities and bio-events.

Although this is a major improvement, the underlying event extraction only considers the occurrence of bio-events (i.e., it only identifies *event-triggers*) and ignores the event participants altogether. This is partly because FACTA+ only uses information on high-level (more abstract) occurrences of concepts. Another limitation of the system is that it does not consider semantically important contextual information, especially modality and negation information about the events. This has an

adverse effect on system performance, as many false-positive associations are retrieved. As mentioned above, these limitations can be addressed by incorporation of relevant contextual information at the event level.

Gene Ontology Term Annotation

The Gene Ontology (GO) [30] provides structured, controlled vocabularies of terms describing gene and gene product characteristics. GO is one of the most commonly used referencing tools in biomedical text mining [2]. Although automated methods have been applied to the task [89], GO annotations of the highest relevance and quality are achieved through manual annotation, by curators reading full-text papers. However, the creation of manual GO annotations is highly expensive. To reduce these annotation costs, significant effort has been focused on the development of systems for automatic annotation and annotator support, and it has been shown that event annotation can assist in the automatic derivation of GO annotations [44].

2.4 Interpretation of Bio-events

Although bio-event extraction has received significant attention in the last few years, identification of contextual information necessary for the correct interpretation of bio-events has been an understudied area. Furthermore, since the majority of research in bio-event extraction has been focussed on the datasets provided by the BioNLP shared tasks, most efforts to recognise contextual interpretative information about events have also been mainly limited to the types of annotations provided in these corpora, i.e., those pertaining to negation and speculation. Moreover, since the recognition of this information was optional in the BioNLP shared tasks, there are

only a fairly small number of event extraction systems that can recognise even these limited types of interpretative information.

Although making such distinctions between events is undoubtedly important, restricting information about event interpretation to these two dimensions (i.e., polarity and certainty level) is often not sufficient to capture the many distinctions between the interpretations of events, which can be both subtle and significant. For instance, the BioNLP shared tasks make a simple binary distinction between speculated and non-speculated events. However, speculation can be expressed to varying degrees, and the ability to distinguish between these could be useful for certain tasks, e.g., slight hedging indicates that the authors are quite confident about the results of their analyses, but they may include a hedging device as a safeguard. In contrast, larger amounts of speculation can indicate that the event should be considered as a hypothesis.

Events that do not have an explicit specification of speculation may nonetheless have different interpretations. An event may be presented as the subject of an investigation, a known fact, an experimental observation or as an outcome of analysing experimental results. A further potential distinction is between events that represent knowledge cited from a previously published paper and events that constitute part of the new knowledge contribution in the paper under consideration. Depending on the nature and criticality of the task being undertaken, some or all of the above distinctions may be important when searching for events in text.

For certain tasks and users, only events that are presented as being completely factual and definite may be sufficient. In other cases, users may be interested specifically in locating events that constitute new experimental knowledge. Tasks where the location of new knowledge is important include building and updating models of bio-

logical processes, such as pathways and curation of biological databases. Such resources are updated by searching the literature for information that can help to enhance and build upon existing, but incomplete, models of a biological process [90]. In order to ensure that such resources are kept as reliable as possible, more tentative results or hypotheses should also be excluded. In the case that the event is presented as an analytical conclusion, it may be important to find appropriate evidence that supports this claim [91] before allowing it to be added to the database.

Other users may be interested in checking for inconsistencies or contradictions in the literature. As an example, consider a case in which two events with identical participants and ontological types appear in two different articles, but one is stated as being positive, whilst the other is negative. If the textual context of both events shows them to have been stated as facts, then this could constitute a serious contradiction. If, however, one of the events is marked as being a hypothesis, then the consequences are not so serious, since the hypothesis may have been later (within the article) reported as being disproved. These issues are further discussed in section 3.1.

Chapter 3: Meta-knowledge

This chapter provides an overview of the design and development of the meta-knowledge annotation scheme. The chapter begins with a discussion on why meta-knowledge annotation is required for bio-events. This is followed by an analysis of related work. Subsequently, a detailed description of the annotation scheme along with a set of hypothetical annotation examples is provided.

3.1 Need for Meta-knowledge Annotation

As explained in chapter 2, recent research in bio-event annotation and extraction has allowed the creation of event-based information retrieval systems with increased power and more focussed searching. However, typical event annotations do not capture contextual information from the sentence, which can be vital for the correct interpretation of the event [91]. Let us consider again sentence S1:

S1 *These results suggest that p21ras proteins are activated by IL-2 in normal human T lymphocytes.*

The phrase at the beginning of the sentence (i.e., *The results suggest that*…) allows us to determine the following about the event that follows:

- It is based on an analysis of experimental results
- It is stated with a certain amount of speculation (evidenced by the use of the verb *suggest,* rather than a more definite verb, such as *demonstrate*).

Altering the words in the context of the event can affect its interpretation in both subtle and significant ways. Consider the hypothetical examples below. Note that

the event-triggers have been underlined and the words/phrases expressing interpretative information have been emboldened:

S3 *It is **known** that the narL gene product activates the nitrate reductase operon*

S4 *We **examined** whether the narL gene product activates the nitrate reductase operon*

S5 *The narL gene product did **not** activate the nitrate reductase operon*

S6 *These results suggest that the narL gene product **might** be activated by the nitrate reductase operon*

S7 *The narL gene product **partially** activated the nitrate reductase operon*

S8 ***Previous studies** have shown that the narL gene product activates the nitrate reductase operon*

If only the event type and participants are considered, then the events in all of the above sentences (S3-S8) are identical to the event in sentence S1. Therefore, a typical event extraction system will interpret all of the above sentences in the same manner, i.e., it will extract the same *Positive Regulation* event from all of the above sentences. However, it is obvious that the knowledge being conveyed in each of the above sentences is significantly different from the others. In sentence S3, the word *known* tells us that the event is a generally accepted fact. However, in sentence S4, the interpretation is completely different. The word *examined* shows that the event is under investigation, and hence the truth value of the event is unknown. The presence of the word *not* in sentence S5 shows that the event is negated, i.e. it did not happen. In sentence S6, the presence of the word *might* (in addition to *suggest*) adds further

speculation regarding the truth of the event. The word *partially* in S7 does not challenge the truth of the event, but rather conveys the information that the strength or intensity of the event is less than what may be expected by default. The phrase *previous studies* in S8 shows that the event is based on information presented in previously published studies, rather than relating to new information from the current study.

Therefore, it is important to consider the context in which the event occurs, since a wide range of different types of information may be expressed that relate directly to the interpretation of the event. We use the term *meta-knowledge* to collectively refer to the different types of interpretative information available in the above sentences.

There are several tasks in which biologists have to search and review the literature that could benefit from the automatic recognition of meta-knowledge about events. These tasks include building and updating models of biological processes, such as pathways [10], and curation of biological databases [30, 92]. Central to both of these tasks is the identification of *new knowledge* that can enhance these resources, e.g., to build upon an existing, but incomplete model of a biological process [90] or to ensure that the database is kept up to date. New knowledge should correspond to experimental findings or conclusions that relate to the current study, which are stated with a high degree of confidence, rather than, e.g., more tentative hypotheses. In the case of an analytical conclusion, it may be important to find appropriate evidence that supports this claim before allowing it to be added to the database [91].

Other users may be interested in checking for inconsistencies or contradictions in the literature. The identification of meta-knowledge could also help to flag such information. Consider, for example, the case where an event with the same ontological

type and identical participants is stated as being true in one article and false in another. If the textual context of both events shows them to have been stated as facts, then this could constitute a serious contradiction. If, however, one of the events is marked as being a hypothesis, then the consequences are not so serious, since the hypothesis may have been later disproved. The automatic identification of meta-knowledge about events can clearly be an asset in such scenarios, and can prevent users from spending time manually examining the textual context of each and every event that has been extracted from a large document collection in order to determine the intended interpretation.

In response to the issues outlined above, we developed a new annotation scheme that is specifically tailored to enriching biomedical event corpora with meta-knowledge, in order to facilitate the training of more useful systems in the context of various information extraction tasks performed on biomedical literature. Our scheme has been designed to be as portable as possible, in that it is not tied to a particular event annotation scheme. This allows the scheme to be applied to a variety of existing event corpora, which generally employ different annotation schemes.

As illustrated by the example sentences above, a number of different types of meta-knowledge may be encoded in the context of an event, e.g., general information type (fact, experimental result, analysis of results), level of confidence/certainty towards the event, polarity of the event (positive or negative), etc. In order to account for this, our annotation scheme is multi-dimensional, with each dimension encoding a different type of information. Each of the 5 dimensions has a fixed set of possible values. For each event, the annotation task consists of determining the most appro-

priate value for each dimension. Textual cue expressions that are used to determine the values are also annotated, when they are present.

3.2 Analysis of Related Work

Although our approach to annotating multi-dimensional meta-knowledge information at the level of events is novel, the more general study of how knowledge in biomedical texts can be classified to aid in its interpretation is a well-established research topic. Two main threads of research can be identified, i.e.:

1) Construction of classified inventories of lexical markers (i.e., words or phrases) which can accompany statements to indicate their intended interpretation.

2) Production of corpora annotated with various different types of meta-knowledge at differing levels of granularity.

3.2.1 Lexical Markers of Meta-Knowledge

The presence of specific cue words and phrases has been shown to be an important factor in classifying biomedical sentences automatically according to whether or not they express speculation [93, 94]. Corpus-based studies of hedging (i.e. speculative statements) in biological texts [95, 96] reinforce the above experimental findings, in that 85% of hedges were found to be conveyed lexically, i.e., through the use of particular words and phrases, rather than through more complex means, e.g., by using conditional clauses. The lexical means of hedging in biological texts have also been found to be quite different to academic writing in general, with modal auxiliaries (e.g., *may, could, would*, etc.) playing a more minor role, and other verbs, adjectives and adverbs playing a more significant role [95]. It has also been shown that, in ad-

dition to speculation, specific lexical markers can denote other types of information pertinent to meta-knowledge identification, e.g., markers of certainty [97], as well as deductions or sensory (i.e. visual) evidence [95].

Based on the above, we can conclude that lexical markers play an important role in distinguishing several different types of meta-knowledge, and also that there are a potentially wide range of different markers that can be used. For example, [98] identified 190 hedging cues that are used in biomedical research articles. Previous work [99] on identifying and categorising lexical markers of meta-knowledge demonstrated that such markers are to some extent domain-dependent. In contrast to other studies, we took a multi-dimensional approach to the categorisation, acknowledging that different types of meta-knowledge may be expressed through different words in the same sentence. As an example, consider sentence S9.

S9 *The DNA-binding properties of mutations at positions 849 and 668* ***may indicate*** *that the catalytic role of these side chains is associated with their interaction with the DNA substrate.*

Firstly, the word *indicate* denotes that the statement following *that* is to be interpreted as an analysis based on the evidence given at the beginning of the sentence (rather than, e.g., a well-known fact or a direct experimental observation). Secondly, the word *may* conveys the fact that the author only has a medium level of confidence regarding this analysis.

Although such examples serve to demonstrate that a multi-dimensional approach recognising meta-knowledge information is necessary to correctly capture potential nuances of interpretation, it is important to note that taking a purely lexical approach to recognising meta-knowledge is not sufficient (i.e., simply looking for words from

these lists that co-occur in the same sentences as events of interest). The reasons for this include:

1) The presence of a particular marker does not guarantee that the "expected" interpretation can be assumed [100]. Some markers may have senses which vary according to their context. As noted in [101], "Every instance should ... be studied in its sentential context" (p.125).

2) Although lexical markers are an important part of meta-knowledge recognition, there are other ways in which meta-knowledge can be expressed. This has been demonstrated in a study involving the annotation of rhetorical zones in biology papers (e.g., background, method, result, implication, etc.) [102], based on a scheme originally proposed in [103]. An analysis of features used to determine different types of zone in the biology papers revealed that in addition to explicit lexical markers, features such as the main verb in the clause, tense, section, position of the sentence within the paragraph and presence of citations in the sentence can also be important.

Thus, rather than assigning meta-knowledge based only on categorised list of cue words and expressions, there is a need to produce corpora annotated with meta-knowledge, on which enhanced information extraction systems can be trained. By annotating meta-knowledge information for each relevant instance (e.g., an event), regardless of the presence of particular lexical markers, systems can be trained to use other types of features that can help to assign meta-knowledge values. However, given that the importance of lexical markers in the recognition of meta-knowledge has been clearly illustrated, we believe that explicit annotation of such markers should be carried out as part of the annotation process, whenever they are present.

3.2.2 Existing Corpora with Meta-Knowledge Annotations

There are several existing corpora with some degree of meta-knowledge annotation. These corpora vary in both the richness of the annotation added, and the type / size of the units at which the meta-knowledge annotation has been performed. Taking the unit of annotation into account, we can distinguish between annotations that apply to continuous text spans, and annotations that have been performed at the event level.

3.2.2.1 Corpora with Meta-Knowledge Annotations at the Text Span Level

Annotations applied to continuous text spans most often only cover a single aspect of meta-knowledge, and are most often carried out at the level of the sentence. The most common types of meta-knowledge annotated correspond to either speculation/certainty level, e.g., [93, 94] or general information content/rhetorical intent, e.g., background, methods, results, insights, etc. This latter type of annotation has been attempted both on abstracts [104, 105] and full papers [102, 103, 106], using schemes of varying complexity, ranging from 4 categories for abstracts, up to 14 categories for one of the full paper schemes. A few schemes annotate more than one aspect of meta-knowledge. For example, [107] annotates both speculation and negation, together with their scopes. Uniquely amongst the corpora mentioned above, [107] also annotates the cue expressions (i.e., the negative and speculative keywords) on which the annotations are based.

Although sentences or larger zones of text [103] constitute straightforward and easily identifiable units of text on which to perform annotation, a problem is that a single sentence may express several different pieces of information, as illustrated by sentence S10.

S10 *Inhibition of the MAP kinase cascade with PD98059, a specific inhibitor of MAPK kinase 1,* ***may*** *prevent the rapid expression of the alpha2 integrin subunit.*

This sentence contains at least 3 distinct pieces of information:

- Description of an experimental method: *Inhibition of the MAP kinase cascade with PD98059*
- A general fact: *PD98059 is a specific inhibitor of MAPK kinase 1*.
- A speculative analysis: *Inhibition of the MAP kinase may prevent the expression of the alpha2 integrin subunit*

The main verb in the sentence (i.e., *prevent*) is the trigger of the speculated event. In a sentence-based annotation scheme, this is likely to be the only information that is encoded. However, this means that other potentially important information in the sentence is disregarded. Some annotation schemes have attempted to overcome such problems by annotating meta-knowledge below the sentence level, i.e., clauses [108, 109] or segments [110]. In the case of the latter scheme, a new segment is created whenever there is a change in the meta-knowledge being expressed. The scheme proposed for segments is more complex than the sentence-based schemes in that it covers multiple types of meta-knowledge, i.e., focus (content type), polarity, certainty, type of evidence and direction/trend (either increase or decrease in quantity/quality). It has, however, been shown that training a system to automatically annotate along these different dimensions is highly feasible [111].

3.2.2.2 Corpora with Meta-Knowledge Annotations at the Event Level

At the level of biomedical events, annotation of meta-knowledge is generally very basic, and is normally limited to negation, e.g., [112]. Negation is also the only attribute annotated in the corpus described in [113], even though a more complex scheme involving certainty, manner and direction was also initially proposed. To our knowledge, only the GENIA Event corpus [29] goes beyond negation annotation, in that different levels of certainty (i.e. *probable* and *doubtful*) are also annotated.

Despite this current paucity of meta-knowledge annotation for events, our earlier examples have demonstrated that further information can usefully be identified at this level, including at least the general information content of the event, e.g. fact, experimental observation, analysis, etc. A possibility would be to "inherit" this information from a system trained to assign such information at the text span level (e.g. sentences or fragments), although this would not provide an optimal solution. The problem lies in the fact that text spans constitute continuous stretches of text, but events do not. The different constituents of an event annotation (i.e., trigger and participants) can be drawn from multiple, discontinuous parts of a sentence. There are almost always multiple events within a sentence, and different constituents of events may be drawn from multiple sentence fragments. This means that mapping between text span meta-knowledge to event-level meta-knowledge cannot be carried out in a straightforward manner. Thus, for the purposes of training more sophisticated event-based information search systems, annotation of meta-knowledge directly at the event level can provide more precise and accurate information that relates directly to the event.

3.3 Meta-knowledge Annotation Scheme

Based on the above analysis, we embarked upon the design of an event-based meta-knowledge annotation scheme specifically tailored for biomedical events. The aim of our meta-knowledge scheme was to capture as much useful information as possible that is specified about individual events in their textual context, in order to support users of event-based search systems in a number of tasks, including the discovery of new knowledge and the detection of contradictions. In order to achieve this aim, our annotation scheme identifies 5 different dimensions of information for each event, taking inspiration from previous multi-dimensional schemes (e.g. [110, 113]). In addition to allowing several distinct types of information to be encoded about events, a multi-dimensional scheme is advantageous, in that the interplay between the different dimension values can be used to derive further useful information (*hyper-dimensions)* regarding the interpretation of the event.

Each dimension of the meta-knowledge scheme consists of a set of complete and mutually-exclusive categories, i.e., any given bio-event belongs to exactly one category in each dimension. The set of possible values for each dimension was determined through a detailed study of over 100 event-annotated biomedical abstracts. In order to minimise the annotation burden, the number of possible categories within each dimension has been kept as small as possible, whilst still respecting important distinctions in meta-knowledge that have been observed during our corpus study. Due to the demonstrated importance of lexical cues in the identification of certain meta-knowledge categories, the annotation task included identification of such cues, when they are present.

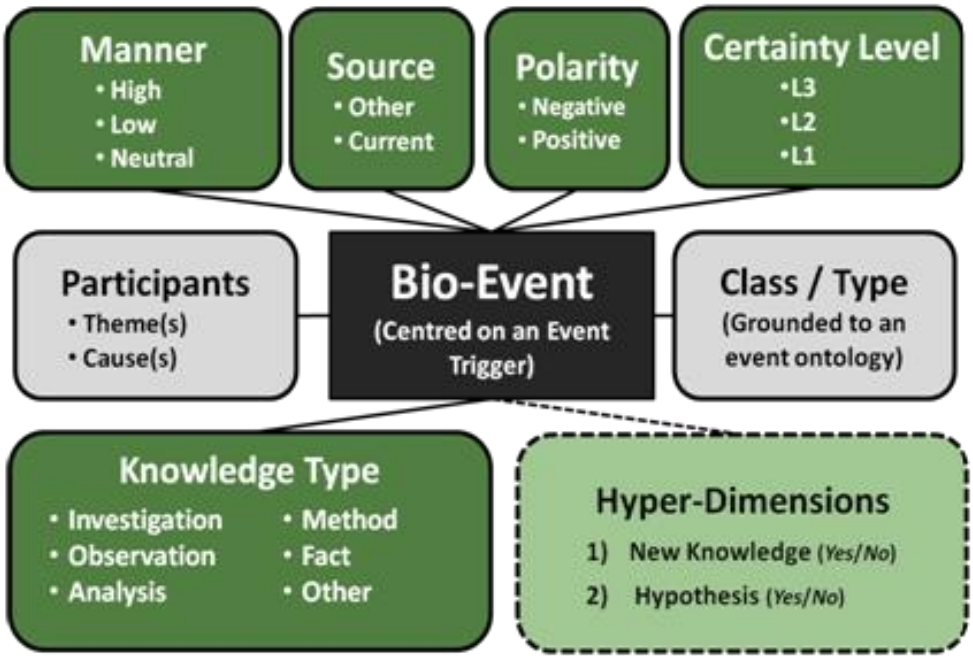

Figure 4. Meta-knowledge Annotation Scheme

Figure 4 provides an overview of the annotation scheme. The boxes with the grey background correspond to information that is common to most bio-event annotation schemes, i.e., the participants in the event, together with an indication of the class or type of the event. The boxes with the dark green backgrounds correspond to our proposed meta-knowledge annotation dimensions and their possible values, whilst the light green box (with a dotted outline) shows the hyper-dimensions that can be derived by considering a combination of the annotated dimensions. Below, we provide a description of each annotation dimension. Further details and examples are provided in the comprehensive (65-page) annotation guidelines, which are provided as an appendix to this thesis.

3.3.1 Knowledge Type

This dimension is responsible for capturing the general information content of the event. The type of information encoded is at a slightly different level to some of the

comparable sentence-based schemes, which have categories relating to structure or "zones" within a document, e.g. *Background* or *Conclusion.* Rather, our *Knowledge Type* dimension attempts to identify a small number of more general information types that can be used to characterise events, regardless of the zone in which they occur. As such, our scheme can be seen as complementary to structure or zone-based schemes, providing a finer-grained analysis of the different types of information that can occur within a particular zone. Our annotation scheme identifies the following 6 categories for this dimension.

3.3.1.1 Investigation

This category is assigned to events indicating enquiries or investigations, which have either already been conducted or are planned for the future.

Evidence

Investigation events are always denoted through an explicit word or phrase in the same sentence as event. Typical types of evidence include:

- Investigative verbs in finite form (i.e., showing tense), e.g., *examine, investigate, analyze/analyse, evaluate, study, test, compare, focus* and *explore* etc. Examples S11-S14 (below) correspond to such cases. These *Investigation* cues normally precede the event-trigger, as in S11 - S13. However, in the case of passive sentences (e.g. S14), the cue appears after the event-trigger.
- Nominalisations of the above verbs (e.g. *investigation, examination, analysis,* etc.) can also indicate investigations, e.g., in S15.
- Verbs in the infinitive form (i.e., preceded by *to)*. These normally precede the event-trigger. The verbs that may be used include all of the above, along with

some others like *define, ascertain, identify* and *elucidate,* etc. An example is shown in S16.

Typical Position in Text

In abstracts, these events typically appear in the beginning of the text, and describe the main investigation(s) reported in the article.

Example Sentences

The following sentences show examples of *Investigation* events:

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

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

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

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

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

S16 *In order to **define** the roles of these two factors, which bind to the same kappa B enhancers, in transcription activation we have prepared somat-*

ic cell hybrids between IARC 301.5 and a murine myeloma.

3.3.1.2 Observation

This category is assigned to events indicating direct observations.

Evidence

Typical evidence for *Observation* events is:

- An explicit cue in the same sentence. Typical cues are verbs like *find, detect* and *observe,* etc. (S17-S19)
- If explicit cues are not present, the event-trigger verb may provide evidence for the assignment of the *Observation* category, if it is:
    - in the past tense (S20-S21)
    - in the present tense, and in an appropriate context (S22)
- Events in document titles (S23)

Typical Position in Text

- Towards the middle of the abstract, following descriptions of background facts and knowledge, and descriptions of investigations to be carried out, but before analyses of results.
- Titles tend to describe definite experimental outcomes and results, unless there is any suggestion to the contrary.

Example sentences

S17 *It was **found** that lipopolysaccharide induced strongly both c-fos and c-jun expression as well as AP1 formation.*

S18 *However, no loss of DNA binding activity is **observed**, presumably reflecting the unique C-terminal domain that is distinct from that present in NF-kappa B p65.*

S19 *Constitutive DNA binding activity consisting of p50 homodimers was **detected** in nuclear extracts from both cell types.*

S20 *LTB4 **increased** the expression of the c-fos gene in a time- and concentration-dependent manner.*

S21 *Both messages rapidly **declined** thereafter.*

S22 *U937 cells **express** both type I and type II IFN receptors.*

S23 *Leukotriene B4 **stimulates** c-fos and c-jun gene transcription and AP-1 binding activity in human monocytes.*

Discussion

If the sentence is in the present tense and an explicit *Observation* cue is not present in the sentence, then the context of the sentence becomes more important in determining the *Knowledge Type* value of the event. For example, consider the case in sentence S22: Taken in isolation, the *Gene Expression* event (centred on the word *express*) seems to be a general scientific fact. However, when we consider that the sentence S24 (below) precedes S22, it transpires that the event in S22 is actually describing an observation.

S24 We have **found** that ISG *expression* in the monocytic U937 cell line differs from most cell lines previously examined.

Taking account of the position of the sentence within the text is often key to determining the correct *Knowledge Type* category. The following two points indicate gen-

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

1) Events occurring in the present tense towards the beginning of an abstract are most likely to correspond to factual statements (i.e., *Knowledge Type* value of *Fact*), unless the context changes this interpretation.

2) Some abstracts are written completely in the present tense. In this case, there is normally an explicit boundary between background knowledge and observations/results. This normally takes the form of a sentence containing an explicit *Observation* cue. The observation interpretation is then normally understood to be "projected" onto events in sentences that follow, that are otherwise unmarked with *Observation* cues. For example, consider the sentence S24. The presence of the word *found* explicitly indicates that an observation is being described. Sentences that follow but are not explicitly marked with cues are highly likely also to describe observations.

Finally, sentence S23 corresponds to an abstract title. Because of this, it can be assumed that the event centred on the verb *stimulates* is describing new knowledge which has been discovered during the study reported in the paper, and hence the event is assigned the *Observation* category.

3.3.1.3 Analysis

This category is assigned to events describing inferences, interpretations, hypotheses or other types of cognitive analysis.

Evidence

Analysis events are always identifiable through the presence of an explicit cue (word or phrase). Typical evidence includes:

- Analysis verbs (finite forms) or their nominalisations preceding the event-trigger, for example, *show, demonstrate, believe, hypothesize, suggest, indicate, appear, seem, conclude, evidence, assume, presume, identify, define, establish, report* and*, reveal*, etc. (S25-S28)
- Conjunctions such as *therefore* and *thus*, etc. These words provide a link to the previous sentence, and imply that some kind of analysis of the results stated in the previous sentence has been carried out in order arrive at the stated event. (S29-S30)
- Verbs or nominalisations serving as event-triggers, for example, *correlate, associate, relate, due to, implicate, attribute, result*, etc. (S31-S32)
- Modal auxiliaries like *may, might* and *could*, as well as adverbs/adjectives like *probably/probable, likely* and *perhaps*. These indicate an uncertainty on the part of the author. As such, they also act as markers of the *Certainty Level* dimension (see section 3.3.2). As this uncertainty must have been reached through some kind of cognitive analysis, they can act as *Analysis* cues, but only when no other *Analysis* cues are present in the sentence, e.g., in S33-S34.
- Frequency indicators such as *often, frequently, normally* and *occasionally* (again, when no other *Analysis* cues are present in the sentence). These de-

note an analysis on the part of the author as to the perceived frequency of occurrence of the specified event. (S36-S37)

- Adjectives and adverbs (mostly non-finite verb forms) like *is able to, is capable of, suggestive of, consistent with, judged by* and *potential,* etc. These again denote analyses on the part of the author. (S38-39)

Typical Position in Text

Towards the end of the abstract, constituting analyses/interpretations of observations and results described previously.

Example Sentences

S25 *These results **indicate** that LTB4 may regulate the production of different cytokines by modulating the yield and/or the function of transcription factors such as AP-1-binding proto-oncogene products.*

S26 *The data **suggest** that differences in functional responses elicited in monocytes by all three factors may be dependent on different routes on nuclear signaling employed by the factors.*

S27 *Unexpectedly, our in vivo studies also **demonstrate** that I kappa B/MAD-3 binds directly to NF-kappa B p50.*

S28 *We also present **evidence** that IL-6 kappa B binding factor II functions as a repressor specific for IL-6 kappa B-related kappa B motifs in lymphoid cells.*

S29 ***Therefore**, an indirect interaction occurs between these two sites*

S30 ***Thus**, both NF-kappa B-binding complexes are needed for optimal viral*

transcription.

S31 *Together, this evidence strongly **implicates** BSAP in the regulation of the CD19 gene.*

S32 *Moreover, in human T helper (Th) clones functionally characterized as being of the type 0, type 1 and type 2 (28%, < 1% und 93% CD30+, respectively), the extent of CD30-mediated NF-kappa B activation **correlated** with the proportion of CD30+ cells.*

S33 *They bind to the kappa B motifs with different relative affinities that **may** reflect their different contribution in the expression of various promoters.*

S34 *The MAD-3 cDNA encodes an I kappa B-like protein that is **likely** to be involved in regulation of transcriptional responses to NF-kappa B, including adhesion-dependent pathways of monocyte activation.*

S35 *Taken together, these observations **suggest** that HIV gene expression **may** be activated in infected monocytes through interaction of the cells with complement-opsonized particles.*

S36 *Our studies now demonstrate that HTLV-1 Tax activates the recently identified cellular kinases IkappaB kinase alpha (IKKalpha) and IKKbeta, which **normally** phosphorylate IkappaB alpha on both of its N-terminal regulatory serines in response to tumor necrosis factor alpha (TNF-alpha) and interleukin-1 (IL-1) stimulation.*

S37 *The activation of transcriptional factor c-Fos/c-Jun AP-1 is essential for normal T cell responsiveness and is **often** impaired in T cells during ag-*

ing.

S38 *In addition, IL-2* ***is capable of*** *increasing transcript levels of the p50 gene coding for the p50 subunit of the NF-kappa B transcription factor, whereas mRNA levels of the p65 NF-kappa B gene remained unchanged.*

S39 *This increase in p50 homodimers coincides with an increase in p105 mRNA,* ***suggestive of*** *a transcriptional up-regulation of p50.*

3.3.1.4 Method

This category is assigned to events that describe experimental methods.

Evidence

Typical evidence is in the form of event-triggers that describe experimental methods, e.g., words like *stimulate, stimulation, addition, pretreated* and *incubated,* etc. (S40-S41)

Typical Position in Text

These events are normally found in the middle part of abstracts, where experimental methods are described.

Example Sentences

S40 *Deoxycholate treatment of the cytoplasmic extract prepared from cells* ***stimulated*** *by TNF-alpha in the presence of Cu2+ resulted in the release of NF kappa B from I kappa B alpha, indicating that Cu2+ interferes with the dissociation of the NF kappa B-I kappa B complex.*

S41 *In addition,* ***pretreatment*** *of the cells with the proteasome inhibitor N-Ac-Leu-Leu-norleucinal inhibits this ligand-induced degradation and, in*

agreement with previous studies, stabilizes a hyperphosphorylated form of the human I kappa B alpha protein.

3.3.1.5 Fact

This category is assigned to events describing general facts and well established knowledge.

Evidence

- Events with triggers that describe biological processes in the present tense. (S42-S43)
- Events contained within relative clauses (S43)
- Explicit cues, such as *known*. (S44)

Typical Position in the Text

Events of this category normally appear towards the beginning of the text, describing background knowledge.

Example Sentences

S42 *Leukotriene B4* ***stimulates*** *c-fos and c-jun gene transcription and AP-1 binding activity in human monocytes.*

S43 *The c-jun mRNA, which is constitutively* ***expressed*** *in human peripheral-blood monocytes at relatively high levels, was also slightly* ***augmented*** *by LTB4*

S44 *Oxidants such as hydrogen peroxide are* ***known*** *to activate certain transcription factors such as nuclear transcription factor kappa beta.*

Discussion

When the main event in a sentence or clause corresponds to an observation, *Fact* events can still occur, e.g. to give further factual information which is necessary to fully explain the event. For example, in S43 the main event of the sentence is centred on *augmented* and is an observation. However, the event centred on *expressed* is providing additional, factual information.

3.3.1.6 Other

This is the default category, which is assigned to events that either do not fit into one of the above categories, do not express complete information, or whose knowledge type is unclear or is assignable from the context. These are mostly *non-propositional* events, i.e., events which cannot be ascribed a truth value due to lack of available (contextual) information.

Evidence

- Secondary (i.e., non-propositional) events whose primary event has the *Knowledge Type* value of *Analysis, Investigation* or *Fact*. (S45-S46)
- Secondary events whose primary event has been negated (i.e., *Polarity = Negative*). (S47)
- Secondary events whose primary event has the *Knowledge Type* value of *Observation*, where the meaning of the trigger verb of the primary event conveys the fact that the secondary event did not take place. Examples of such cue words include *inhibit* and *suppress,* etc. (S48)
- Events that describe properties of entities. (S49).

Typical Position in Text

These events do not have a typical position within text and are usually scattered over the entire abstract.

Example Sentences

S45 *These results **indicate** that LTB4 may regulate the production of different cytokines.*

S46 *The effects of prostaglandin E2 (PGE2) on cytokine production and proliferation of the CD4+ human helper T cell clone SP-B21 were **investigated**.*

S47 *Integrin ligation with antibodies does not induce tyrosine phosphorylation of FAK.*

S48 *In vitro translated MAD-3 protein was **found** to specifically inhibit the DNA-binding activity of the p50/p65 NF-kappa B complex*

S49 *A Rel-related, mitogen-inducible, kappa B-binding protein has been cloned as an immediate-early activation gene of human peripheral blood T cells.*

Discussion of Examples

In (S45) the primary event, whose trigger is *regulate*, is an *Analysis* event, according to the presence of the word *indicate.* However, there is a secondary event whose trigger is *production.* The analysis interpretation does not extend to this secondary event, i.e. the interpretation of this event is not that "production of different cytokines *may* occur". In fact, the secondary event does not have a specific interpretation, e.g. there is nothing providing information about whether it is a general fact or

under what circumstances it occurs. In other words, it has an incomplete interpretation when considered in isolation from the primary event. For this reason, it would be assigned the *Knowledge Type* value of *Other*. Sentence (S46) shows a similar case, where the primary event, whose trigger is *effects,* has the *Knowledge Type* value of *Investigation.* The secondary events whose triggers are *production* and *proliferation* would thus be assigned the *Knowledge Type* value of *Other*.

In (S47), the fact that the primary event (whose trigger is *induce*) is negated, means that the secondary event (with trigger *phosphorylation*) did not take place. The primary event is an *Observation* (according to the context in which it appears). However, the secondary event was not observed, and hence it would be assigned a *Knowledge Type* value of *Other*.

Sentence (S48) exhibits a similar behaviour. The primary event has the trigger *inhibit.* Although this is an *Observation* (based on the presence of the word *found*), the negative meaning of *inhibit* means that the secondary binding event did not take place. Therefore, the secondary event would be assigned the *Knowledge Type* value of *Other*.

In (S49), the *Positive Regulation* event centred on *inducible* describes a property of the protein, namely that it is induced by mitogen.

3.3.2 Certainty Level

This dimension aims to identify the level of certainty associated with the occurrence of the event, as ascribed by the authors. It comes into play whenever there is an explicit indication that there is less than complete confidence that the specified event will occur. This could be because:

- There is uncertainty regarding the general truth value ascribed to the event.
- It is perceived that the event may not take place all of the time.

Different degrees of uncertainty and frequency can be considered as points on a continuous scale, and there is an on-going discussion regarding whether it is possible to partition the epistemic scale into discrete categories [114]. However, the use of a number of distinct categories is undoubtedly easier for annotation purposes and has been proposed in a number of previous schemes. Although recent work has suggested the use of four or more categories [99, 111, 114], our initial analysis of bio-event corpora showed that only three levels of certainty seem readily distinguishable for bio-events. This is in line with [115], whose analysis of general English showed that there are at least three articulated points on the epistemic scale.

Like [110], we have chosen to use numerical values for the Certainty Level dimension, in order to reduce potential annotator confusions or biases that may be introduced through the use of labels corresponding to particular lexical markers of each category, such as *probable* or *possible*. Such labels could in any case be misleading, given that frequency can also come into play in assigning the correct category. Our chosen values of the Certainty Level dimension are defined as follows:

3.3.2.1 L3

This is the default category. No explicit expression that either:

- there is uncertainty or speculation towards the event
- the event does not occur all of the time

3.3.2.2 L2

Explicit indication of either:

- High (but not complete) confidence or slight speculation towards the event.
- The event occurs frequently, but not all of the time.

Evidence

These events are always indicated through an explicit word or phrase in the same sentence as event. Typical cues are:

- Words such as *likely* and *probably* (S50-S51).
- Verbs that are also used as cues for the assignment of the *Analysis* category of *Knowledge Type* dimension, which convey the meaning of a somewhat tentative analysis, e.g. *believe, hypothesize, suggest* and *indicate.*(S52-S53)
- Words such as *normally, often, frequently,* etc. (S54-S55).

Example Sentences

S50 *The loss of conventional responsiveness is **probably** caused by alterations at the level of signalling*

S51 *The MAD-3 cDNA encodes an I kappa B-like protein that is **likely** to be involved in regulation of transcriptional responses to NF-kappa B, including adhesion-dependent pathways of monocyte activation.*

S52 *Recently, investigators have **hypothesized** that CD14-mediated signaling is effected through a receptor-associated tyrosine kinase (TK), suggesting a multicomponent receptor model of LPS signaling.*

S53 *During the course of serious bacterial infections, lipopolysaccharide*

*(LPS) is **believed** to interact with macrophage receptors, resulting in the generation of inflammatory mediators and systemic symptoms including hemodynamic instability and shock.*

S54 *Expression of IL-1alpha by HTLV-I productively infected cells may be important in the hypercalcemia, osteolytic bone lesions, neutrophilia, elevation of C-reactive protein, and fever **frequently** seen in patients with HTLV-I-induced adult T-cell leukemia/lymphoma*

S55 *HIV-1-infected myeloid cells are **often** diminished in their ability to participate in chemotaxis, phagocytosis, and intracellular killing.*

3.3.2.3 L1

Explicit indication of either:

- Low confidence or considerable speculation towards the event.
- The event occurs infrequently or only some of the time.

Evidence

These events are always indicated through an explicit word or phrase in the same sentence as event. Typical cues are:

- Words such as *may*, *might* and *perhaps* (S56-S57)
- Verbs that are also used as cues for the assignment of the *Analysis* category of *Knowledge Type* dimension, which convey the meaning of a highly tentative analysis, e.g. *speculate, suppose* and *suspect*, etc.
- Words such as *sometimes, rarely, scarcely,* etc.

Example Sentences

S56 *These results indicate that LTB4* ***may*** *regulate the production of different cytokines by modulating the yield and/or the function of transcription factors such as AP-1-binding proto-oncogene products.*

S57 ***Perhaps*** *murine thymocytes are denied this form of rescue because they shut off IL-2R beta chain expression at an earlier stage*

3.3.3 Polarity

This dimension has been designed to capture the truth value of the assertion encapsulated by the event. We define a negated event as "an event which describes the absence or non-existence of an entity or a process". That is to say, the event may describe that a process does not or did not happen, or that an entity is absent or does not exist. The recognition of such information is vital, as the interpretation of a negated event instance is completely opposite to the interpretation of a non-negated (positive) instance of the same event. Our scheme permits the following two values for this dimension:

3.3.3.1 Positive

No explicit negation of the event (default)

3.3.3.2 Negative

The event has been negated according to the description above.

Evidence

Negated events are always indicated through an explicit word or phrase in the same sentence as event. Typical indicators are:

- The most common means of expressing negation is through the use of the words *not* or *no* (S58-S59).
- A number of other words can also be used to express the fact that an event did not take place, when occurring in certain contexts. Examples include *fail*, *lack*, and *unable*, *exception*, *independent, without* (S60-S62).

Example Sentences:

S58 *CsA was found **not** to inhibit lck gene expression, nor the activity of the lck gene product.*

S59 *Protein synthesis inhibitors and corticosteroids, which suppress arachidonate release and the synthesis of proinflammatory cytokines, had **no** effect on translocation of NF-kappa B in CHO/CD14 or RAW 264.7 cells, demonstrating that NF-kappa B translocation is an early event.*

S60 *In contrast, NF-kappa B p50 alone **fails** to stimulate kappa B-directed transcription, and based on prior in vitro studies, is not directly regulated by I kappa B.*

S61 *The CD19 protein is expressed on the surface of all B-lymphoid cells with the **exception** of terminally differentiated plasma cells*

S62 *Binding of type I interferon (IFN-alpha/beta) to specific receptors results in the rapid transcriptional activation, **independent** of protein synthesis, of IFN-alpha-stimulated genes (ISGs) in human fibroblasts and HeLa and Daudi cell lines.*

Discussion of Examples:

In sentence S61, there are 2 events that are centred on the verb *expressed*. In the first event, the *CD19 protein* is expressed on the surface of *all B-lymphoid cells*, and so the event is positive. In the second event, the presence of the word *exception* denotes the fact that *CD19 protein* is not expressed on *terminally differentiated plasma cells*, and hence this is a negated event.

In example S62, the event centred on the word *independent* denotes an event of type *Correlation* (according to the GENIA Event annotation guidelines) and involves *transcriptional activation* and *protein synthesis.* The use of the word *independent* itself indicates that no correlation exists between them, because the transcriptional activation takes places independently of protein synthesis. Therefore, the correlation event is inherently negative. This example serves to illustrate the potential complexity in recognizing events with negative polarity. Sometimes, the meaning and type of the event have to be considered carefully in order to determine whether it is positive or negative. This issue is discussed in detail in Chapter 5.

3.3.4 Manner

This dimension identifies the rate, level, strength or intensity of the event (in biological terms). Such information has previously been shown to be relevant for biologists. This is evidenced in the event annotation scheme for the GREC corpus [116], which was designed in consultation with biologists, and identified expressions of manner as one of the semantic roles associated with the event. The proposal for the annotation of protein-protein interactions suggested in [113] also lists manner as a potentially useful attribute to annotate. Inspired by these works, we build upon the

types of manner annotation available in the GREC corpus by attempting a three-way categorisation of manner, as explained below:

3.3.4.1 High

This category is assigned to events with explicit indication that the event occurs at a high rate, level, strength or intensity.

Evidence

The events are always indicated through an explicit word or phrase in the same sentence as the event. Typical cues are:

- Adverbs: examples include *strongly*, *rapidly* and *highly*, etc. (S63-S65)
- Adjectives: examples include *high, rapid, profound,* etc. (S66-S68)

Example Sentences

S63 *Both messages **rapidly** declined thereafter.*

S64 *It was found that lipopolysaccharide induced **strongly** both c-fos and c-jun expression.*

S65 *Although IFN-gamma alone does not induce ISG expression, IFN-gamma pretreatment **markedly** increases and hastens ISG expression and transcriptional induction.*

S66 *In particular, the c-Rel homodimer has a **high** affinity for interleukin-6 (IL-6) and beta interferon kappa B sites.*

S67 *However, the **profound** T cell deficit of nude mice indicates that the thymus is by far the most potent site for inducing the expansion per se.*

S68 *Binding of type I interferon (IFN-alpha/beta) to specific receptors results*

*in the **rapid** transcriptional activation.*

Discussion of Examples

Sentence S64 shows a case where *strongly* indicates a high rate of induction. It is important to note that certain words like *strongly* only indicate a high manner when they are modifying verbs that describe biological processes. When used in conjunction with verbs denoting the *Analysis* category of the *Knowledge Type* dimension (e.g. *strongly suggest*), they indicate the *Certainty Level* (rather than the *Manner*) of the event.

In example sentence S65, the manner adverb *markedly* applies both to the events centred on *increases* and *hastens,* to indicate a high level.

3.3.4.2 Low

This category is for events with explicit indication that the event occurs at a low rate, level, strength or intensity.

Evidence

These events are always indicated through an explicit word or phrase in the same sentence as event. Typical cues are:

- Adverbs: examples include *slightly*, *partially*. (S69-S70)
- Adjectives: examples include *little, small, slight.* (S71-S72)
- Phrases such as *barely, scarcely* (*any*)*, almost no.* Although such phrases have negative connotations, they still convey the fact that the stated event took place, even though in a very insignificant way. Hence, the event should have a *Polarity* value of *Positive,* and a *Manner* value of *Low.* (S73-S74)

Example sentences

S69 *The c-jun mRNA was also **slightly** augmented by LTB4.*

S70 *Alteration of the sequence at threonine 78 can **partially** restore function to a verb A protein rendered defective due to a mutation at position 61.*

S71 *Moreover, kappa 1-kappa 3 can each be deleted from the TNF-alpha promoter with **little** effect on the gene's inducibility by PMA.*

S72 *The oxLDL-induced NF-kappa B activation was accompanied by an initial depletion of I kappa B-alpha followed by a **slight** transient increase in the level of this inhibitor protein.*

S73 *In contrast, the RelA(p65) subunit was **barely** detectable in monocytes, but its level increased markedly in MDMs.*

S74 *Tumor necrosis factor induced slightly c-fos and had **almost no** effect on c-jun and AP1.*

3.3.4.3 Neutral

This is the default category. Assigned when there is no explicit indication of either high of low manner, but also in the rare cases when neutral manner is explicitly indicated, using cue words such as *normal* or *medium,* etc. For example, consider the example sentence S75.

S75 *The eukaryotic transcription factor NF-kappa B plays a central role in the induced expression of human immunodeficiency virus type 1 and in many aspects of the genetic program mediating **normal** T-cell activation and growth.*

3.3.5 Knowledge Source

This dimension aims to denote the source or origin of the knowledge being expressed by the event. Specifically, we distinguish between events that can be attributed to the current study, and those that are attributed to other (previous) studies. Information about knowledge source has been demonstrated to be important according to its annotation in both the Gene Ontology [30] and in the corpus presented in [110]. This dimension can help in distinguishing new experimental knowledge from previously reported knowledge. Two possible values are distinguished, as follows:

3.3.5.1 Current

This category is assigned to events that make an assertion that can be attributed to the current study. This is the default category, and is assigned in the absence of explicit lexical or contextual cues, although explicit cues such as *the present study* may be encountered.

Evidence

- Explicit evidence for this category is often not present. Sentences describing results that are unmarked for knowledge source normally correspond to *Current*, although this is not exclusively the case, and context must be examined to determine whether the event refers to the current or a previous study.
- When explicit evidence is present, the word *we* is often present in the sentence. On its own, this is not enough to determine the value of *Current*, as the sentence could be referring to work carried out by the authors in a previous study (see sentence S79 in the discussion below).
- Reliable indicators involving *we* include the following:

- *We have + past participle,* e.g. *we have found that ... .* (S75)
- The use of *here* in conjunction with *we,* e.g. *we report here that ...* denoting that the event is relevant in the current study. (S76)
- Phrases such as *The present work, in this study,* etc. (S77)

Example Sentences:

S75 ***We have*** *examined the effect of leukotriene B4 (LTB4) on the expression of the proto-oncogenes c-jun and c-fos.*

S76 ***We*** *report* ***here*** *that the second alteration, at threonine 78, also plays an important, although more indirect, role.*

S77 ***The present work*** *has examined the effects of okadaic acid, an inhibitor of type 1 and 2A protein phosphatases, on the regulation of c-jun expression during monocytic differentiation of U-937 leukemia cells.*

Discussion of Examples

Consider example S78, which demonstrates how the presence of the word *we* alone is not necessarily sufficient to determine a *Source* value of *Current:*

S78 *In addition,* ***we*** *looked at the modulation of nuclear factors binding specifically to the AP-1 element after LTB4 stimulation.*

In order to determine whether the event marked in S78 has a *Knowledge Source* value of *Current,* the context needs to be examined. In isolation, the use of the simple past tense (*looked at)* is ambiguous as regards the knowledge source, i.e. it may refer to a previous study undertaken by the authors, in which case the *Knowledge Source* value of *Other* would be assigned (see below). Equally, it may refer to the current

study, in which case the *Knowledge Source* value would be set to *Current.* However, S75 and S78 are drawn from the same abstract, where S75 immediately precedes S78. As sentence S75 contains sufficient evidence to link it to the current study, and as sentence S78 is explicitly linked to it through the use of *In addition,* it follows that sentence S78 must also refer to the current study, and hence has a *Knowledge Source* value of *Current.*

Consider the example sentence S79, where no explicit marker of *Knowledge Source* is present in the sentence.

S79 *LTB4* ***increased*** *the expression of the c-fos gene in a time- and concentration-dependent manner.*

Although S79 is fairly clearly an experimental observation, it is only by examining the context that it can be discovered whether this is a result of the current study, or a previous one. At least for abstracts, if a sentence such as S75 occurs towards the beginning of the abstract, then it will normally be the case that any subsequently reported results should be interpreted as being attributable to the *Current* study, unless there is any explicit indication to the contrary.

3.3.5.2 Other

This category is assigned to events that are attributed to a previous study.

Evidence

These events are always indicated through an explicit word or phrase. Typical cues are:

- Words and phrases like *previous studies* and *previously,* etc. (S80-S81).

- Citation of another paper (S82).
- Events that are attributable to the current author, but which implicitly refer to a study other than the current one (S83).

Example sentences

S80 *Although it has been **previously** shown that the IL-6 kappa B motif functions as a potent IL-1/tumor necrosis factor-responsive element in non-lymphoid cells, its activity was found to be repressed in lymphoid cells such as a Jurkat T-cell line.*

S81 *Since **previous studies** have demonstrated that the c-jun gene is autoinduced by Jun/AP-1, we also studied transcription of c-jun promoter (positions -132/+170)-reporter gene constructs with and without a mutated AP-1 element.*

S82 *A recent functional analysis by **Miyatake et al. (S. Miyatake, M. Seiki, M. Yoshida, and K. Arai, Mol. Cell. Biol. 8:5581-5587, 1988)** described a short promoter region in the GM-CSF gene that conferred strong inducibility by T-cell-activating signals and tax1, but no NF-kappa B-binding motifs were identified.*

S83 *We have **earlier** found that in Jurkat cells activation of protein kinase C (PKC) enhances the cyclic adenosine monophosphate (cAMP) accumulation induced by adenosine receptor stimulation or activation of Gs.*

Discussion of Examples

In S83, although the use of the present perfect *we have* would normally indicate that the reported event belongs to the current study, the presence of the word *earlier* shows that event centred on *enhances* is an observation from an earlier study.

3.3.6 Hyper-Dimensions

A defining feature of our annotation scheme is the fact that, in addition to the explicitly annotated dimensions, further information can be inferred by considering combinations of some of these dimensions. We refer to these additional types of information as the *hyper-dimensions* of our scheme, of which we have identified two.

3.3.6.1 New Knowledge

The isolation of events describing new knowledge is, as we have described earlier, important for certain tasks undertaken by biologists. However, it is not possible to determine whether an event represents new knowledge by considering a single annotation dimension. For example, events having been assigned *Knowledge Type=Observation* could correspond to new knowledge, but only if they represent observations from the current study, rather than observations cited from elsewhere. In a similar way, an *Analysis* drawn from experimental results in the current study could be treated as new knowledge, but generally only if it represents a straightforward interpretation of results, rather than something more speculative. Thus, we consider *New Knowledge* to be a hyper-dimension, whose value (either *Yes* or *No*) can be inferred by considering a combination of the value assignments for the *Knowledge Type*, *Source* and *Certainty Level* dimensions.

Table 1 is an inference table that can be used to obtain the appropriate value for *New Knowledge*, based on the values assigned to the three dimensions mentioned above. The symbol 'X' indicates a "don't care condition", meaning that this value does not have any impact on the result.

Source (Annotated)	**Knowledge Type (Annotated)**	**Certainty Level (Annotated)**	**New Knowledge (Inferred)**
Other	X	X	No
X	X	L2	No
X	X	L1	No
Current	Observation	L3	Yes
Current	Analysis	L3	Yes
X	Fact	X	No
X	Method	X	No
X	Other	X	No
X	Investigation	X	No

Table 1. Inference Table for New Knowledge Hyper-Dimension

3.3.6.2 Hypothesis

The second hyper-dimension of our scheme is *Hypothesis*. The binary value of this hyper-dimension can be inferred by considering the values of *Knowledge Type* and *Certainty Level*. Events with a *Knowledge Type* value of *Investigation* can always be assumed to be a hypothesis. However, if the *Knowledge Type* value is *Analysis,* then only those events with a *Certainty Level* value of *L1* or *L2* (speculative inferences made on the basis of results) should be considered as hypothesis, to be matched with

more definite experimental evidence when available. A value of *L3* in this instance would normally be classed as an instance of new knowledge, as indicated in Table 1.

The cases in which an event can be assumed to be a hypothesis are summarised in Table 2.

Knowledge Type (Annotated)	**Certainty Level (Annotated)**	**Hypothesis (Inferred)**
Fact	X	No
Method	X	No
Other	X	No
Observation	X	No
Analysis	L3	No
Analysis	L2	Yes
Analysis	L1	Yes
Investigation	X	Yes

Table 2. Inference Table for Hypothesis Hyper-Dimension

3.4 Hypothetical Annotation Examples

Having examined the different annotation dimensions of the scheme in some detail, we now re-examine the hypothetical sentences first introduced in section 3.1, and discuss the correct categories to assign to them for each meta-knowledge dimension:

S3 *It is **known** that the narL gene product activates the nitrate reductase operon*

Knowledge Type: ***Fact***. The word *known* indicates that this is a generally known fact.

Certainty Level: ***L3***. There are no words or phrases to suggest uncertainty, so the default value of *L3* is assigned.

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

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

Source: ***Current.*** There is no indication of a source other than the current text; hence the default value of *Current* is assigned.

S4 *We **examined** whether the narL gene product activates the nitrate reductase operon*

Knowledge Type: ***Investigation***. The word *examined* indicates that the event describes an investigation.

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

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

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

Source: ***Current.*** There is no indication of a source other than the current text; hence the default value of *Current* is assigned.

S5 *The narL gene product did **not** activate the nitrate reductase operon*

Knowledge Type: ***Observation***. The use of the past tense (along with the lack of any other *Knowledge Type* cues) indicates that this is an experimental observation.

Certainty Level: ***L3***. There are no words or phrases to suggest uncertainty, so the default value of *L3* is assigned.

Polarity: ***Negative.*** The negation cue *not* modifies the event-trigger. This indicates that the event is negated.

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

Source: ***Current.*** There is no indication of a source other than the current text; hence the default value of *Current* is assigned.

S6 *These results suggest that the narL gene product **might** be activated by the nitrate reductase operon*

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

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

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

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

Source: ***Current.*** There is no indication of a source other than the current text; hence the default value of *Current* is assigned.

S7 *The narL gene product **partially** activated the nitrate reductase operon*

Knowledge Type: ***Observation***. The use of the past tense (along with the lack of any other *Knowledge Type* cues) indicates that this is an experimental observation.

Certainty Level: ***L3***. There are no words or phrases to suggest uncertainty, so the default value of *L3* is assigned.

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

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

Source: ***Current.*** There is no indication of a source other than the current text; hence the default value of *Current* is assigned.

S8 ***Previous studies** have shown that the narL gene product activates the nitrate reductase operon*

Knowledge Type: ***Analysis***. The word *shown* is present, indicating that some analysis about the event has been undertaken.

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

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

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

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

Chapter 4: Meta-Knowledge Annotation

This chapter provides an overview of the application of the meta-knowledge annotation scheme to two bio-event corpora. It starts with brief descriptions of the evaluation of the annotation scheme and the training of the annotators. This is followed by detailed discussions about the creation of the two bio-event corpora enriched with meta-knowledge annotations: GENIA-MK (abstracts) and FP-MK (full papers). Finally, a brief comparison between the annotation characteristics of the two corpora is provided.

4.1 Evaluation of the Annotation Scheme

Before embarking on a large scale annotation project, we conducted a small annotation experiment to verify the feasibility and soundness of the meta-knowledge annotation scheme [117]. Two annotators independently applied the annotation scheme to bio-events identified in 70 randomly selected abstracts from the GENIA Pathway corpus [118], using the annotation manual we had developed. The results were encouraging: high rates of inter-annotator agreement (between 0.89 and 0.95 Kappa) were achieved. The experiment helped to demonstrate the soundness of both the scheme itself and the guidelines. Furthermore, the fact that all categories within all dimensions were annotated, at least to a certain extent, suggested that none of the proposed categories was redundant.

4.2 Annotators and Training

In order to ensure the efficacy of the guidelines and the reproducibility of the annotation task, we recruited 2 external annotators to carry out the annotation of a gold

standard corpus. An important consideration was the type of expertise required by the annotators. It has previously been found that at least negations and speculations in biomedical texts can be reliably detected by linguists [107]. The scope of our meta-knowledge annotation is wider, involving some scientifically motivated aspects (i.e., *Knowledge Type* and *Manner*), but the assignment of certain dimension values is somewhat linguistically motivated, e.g., it is often the case that meta-knowledge cue expressions have a grammatical relationship to the event-triggers and participants. In order to verify the extent to which either domain-specific biological knowledge or linguistic knowledge is required to perform the annotation accurately, we recruited a biology expert and a linguistics expert to carry out the task. Both annotators had near-native competency of English, which we considered to be important to carry out the task accurately.

The annotators undertook training prior to commencing the annotation of the gold standard corpus. This training began with initial introductory sessions, in which the annotation scheme and guidelines were explained, and the X-Conc annotation tool [119] was demonstrated. Subsequently, the annotators carried out practice annotation tasks. For this purpose, we used the same corpus of 70 abstracts from the GENIA Pathway corpus that was used to test the feasibility of the scheme, as described above. Both annotators were given the same sets of abstracts to annotate, independently of each other. This allowed us to detect a maximal number of potential annotation errors and discrepancies produced by the annotators, as we could conduct comparisons not only between the annotators themselves, but also against the gold standard annotations which had previously been created. The annotators returned a set of abstracts each week, in response to which we produced detailed

feedback reports highlighting annotation errors. These reports were thoroughly discussed with the annotators, in order to maximally enhance and accelerate the learning process. Sometimes, errors made by the annotators highlighted potential problems with the annotation guidelines, which were addressed by updating the guidelines accordingly.

4.3 Annotation of Abstracts

Following the completion of annotator training, the annotation scheme was applied to enrich the entire GENIA Event corpus [29] with meta-knowledge information. To our knowledge, the enriched corpus, which we refer to as the GENIA-MK corpus, represents a unique effort within the domain, in terms of the amount of meta-knowledge information annotated at such a fine-grained level of granularity (i.e., events). As the GENIA Event corpus is currently the largest biomedical corpus annotated with events, the enrichment of this entire corpus with meta-knowledge annotation constitutes a valuable resource for training information extraction systems to recognise not only the core information about events and their participants, but also additional information to aid in their correct interpretation and to provide enhanced search facilities.

4.3.1 General Corpus Characteristics

In this section, we discuss the general distribution of the annotations amongst the different categories of each dimension, and also provide lists of the most commonly annotated cue expressions.

4.3.1.1 Knowledge Type

Table 3 shows the number of instances of each category annotated for the *Knowledge Type* dimension. The most common category is *Observation,* constituting just over a third of the total number of events. This result is unsurprising, since abstracts would be expected to focus mainly on definite experimental observations and results, both of which fall into this category. The *Other* category is almost as common as *Observation*. Such events are generally the participant events of *Investigation, Analysis* or *Fact* events which, out of the context of their parent event, have no specific *Knowledge Type* interpretation. The total number of *Other* events is very similar to the combined total of *Investigation, Analysis* and *Method* events. This is to be expected, given the high proportion (44%) of complex events present in the corpus.

Category	Frequency	% of total events
Observation	12821	34.7%
Other	11537	31.3%
Analysis	6578	17.8%
Fact	2998	8.1%
Investigation	1948	5.3%
Method	976	2.6%

Table 3. Distribution of annotated categories for *Knowledge Type*

The proportion of *Analysis* events is much smaller but still quite significant, since most abstracts contain at least some analysis of the experimental results obtained. The usual inclusion of a small amount of background factual information to put the current study into context accounts for the average of 3 events per abstract (8% of all

events) that are assigned the *Fact* category. Even briefer are the descriptions of what is to be investigated, with an average of 2 *Investigation* events per abstract (5% of all events). The scarcity of events describing methods (2.6% of events, or less than 1 event per abstract) shows that providing details of experimental setup is very rare within abstracts.

Analysis		**Investigation**		**Observation**	
Cue	**Freq**	**Cue**	**Freq**	**Cue**	**Freq**
suggest	408	examined	207	found	361
show	353	investigated	205	observed	226
demonstrate	335	analysed	119	detected	141
demonstrated	332	studied	94	detectable	48
showed	246	to determine	50	seen	32
shown	244	tested	39	noted	17
may	242	measured	25	find	11
can	232	monitored	25	detect	11
associated	215	to investigate	23	findings	11
indicate	211	to examine	21	observations	9
revealed	196	to study	21	finding	9
suggesting	140	analysis	20	show	6
report	114	studies	20	report	6
identified	112	to identify	16	exhibit	5
thus	108	investigate	15		

Table 4. Most common *Knowledge Type* cue expressions

Table 4 shows the 15 most commonly annotated cue expressions for the *Knowledge Type* categories of *Analysis, Investigation* and *Observation* together with their frequencies. Cues were also annotated for the *Fact* category, if they were present. However, only 139 of the 2998 *Fact* events (4.6%) have a cue expression annotated. Of these annotated cue expressions, 106 (76%) correspond to the word *known*. Cue expression annotation was also optional for the *Observation* category, in which only 937 (7.3%) of the total number of events are accompanied by a cue. For the *Investigation* and *Analysis* categories, all annotated events have a cue expression.

For both *Investigation* and *Observation,* the top three most common cue expressions are past tense verbs, whilst the use of the present tense appears to be more dominant for describing *Analysis* events. The use of infinitive forms (i.e. *to investigate*) as cues seems to be a specific feature of the *Investigation* category. Whilst most cues are verbal forms, words with other parts of speech can sometimes constitute reliable cues (e.g. *thus* for *Analysis,* and *detectable* for *Observation*).

4.3.1.2 Certainty Level

The distribution of *Certainty Level* annotations is shown in Table 5.

Category	Frequency	% of total events
L3 (default)	33876	91.9%
L2	2216	6.0%
L1	766	2.1%

Table 5. Distribution of annotated categories for *Certainty Level*

Despite the relative scarcity of *Certainty Level* marking on events, it should be noted that this dimension is only applicable when the *Knowledge Type* value of *Analysis* is

assigned. Taking this into consideration, the need for this dimension becomes more apparent: whilst over half of *Analysis* events (54.7%) are stated with no uncertainty, this also means that almost a half of these events *do* express some kind of uncertainty. In fact, approximately one third (33.7%) of all *Analysis* events are annotated as *Certainty Level=L2*, whilst 11.6% are reported with less certainty (i.e., *Certainty Level=L1*). The very nature of abstracts means that the high proportion of events with no uncertainty is to be expected. As authors aim to "sell" the most positive aspects of their work in abstracts, it makes sense that the majority of analyses should be presented in a confident manner.

However, the marking of slight uncertainty is sometimes necessary. The author's analyses of experimental results may have produced important outcomes, but yet they are not confident that their analysis is completely reliable. As stated in [95], "Scientists gain credibility by stating the strongest claims they can for their evidence, but they also need to insure against overstatement." (p. 257). Such insurance can often be achieved by the use of slight hedging (i.e., *Certainty Level=L2*). Greater speculation (i.e., *Certainty Level=L1*) is less common, as such credibility is reduced in this case.

As part of the original annotation in the GENIA Event corpus, *Uncertainty* was annotated as an event attribute. The default value is *Certain* and the other two values are *Probable* and *Doubtful.* In the GENIA event annotation guidelines, these attributes do not have clear definitions. However, *Probable* can be defined loosely as something that is hypothesized by the author, while *Doubtful* is something that is investigated. As such, *Probable* has more in common with our *Certainty Level* dimension, while *Doubtful* is more closely linked to the *Investigation* category of our

Knowledge Type dimension. Therefore, the GENIA *Uncertainty* attribute does not distinguish between degrees of uncertainty in the same way as our meta-knowledge scheme. Comparison of results confirms this – of the events annotated with *Uncertainty=Probable,* there are comparable numbers of events that have been annotated with *Certainty Level=L1* (530 events) and *Certainty Level=L2* (665 events). It is also worth noting that the total number of events identified with some degree of uncertainty using our scheme (*Certainty Level=L1* or *Certainty Level=L2*) is 2982. This is almost double the number of events annotated as *Probable*, showing that our more detailed guidelines for *Certainty Level* annotation have helped to identify a far greater number of events expressing some degree of speculation.

Discrepancies can also be found regarding the *Doubtful* category. Whilst, as expected, the vast majority of these correspond to events that have been annotated as *Knowledge Type=Investigation* in our meta-knowledge scheme (1022 out of a total of 1349 *Doubtful* events), some *Doubtful* events also correspond to events with other *Knowledge Type* values (most notably *Analysis* with *Certainty Level* values of *L3, L2* or *L1,* which can also occur within the *Probable* category). This provides evidence that the boundary between *Doubtful* and *Probable* may not always have been clear to annotators. In addition, our scheme identified 1948 events with *Knowledge Type=Investigation*, meaning that there were some 900 investigative events that were not identified during the original GENIA Event annotation.

Table 6 shows the 10 most commonly annotated cue expressions for the *L2* and *L1* values. For *L2*, the most common expression is *can,* which normally expresses ability rather than speculation (together with the cues *ability* and *able*). If an event has the ability to occur, then there is no guarantee that it will occur all of the time, and

hence it is sensible that the event should be annotated as having less than complete certainty.

All of the other words in the *L2* list express slight speculation or hedging, mostly corresponding to different forms of the verbs *suggest* and *indicate.* In Table 4, it was seen that these verbs also rank amongst the most common *Analysis* cues, showing that it is common for analysis and slight speculation to be simultaneously expressed using a single cue word. For the indication of *L1* certainty, modal auxiliary verbs are particularly common, with *may* accounting for 67.4% of all annotated *L1* cues, and *might* and *could* constituting a significant proportion of the remainder. The *L1* category has a very small number of distinct cue expressions (23), compared to 121 distinct expressions for *L2*.

L2		**L1**	
Cue	**Frequency**	**Cue**	**Frequency**
can	407	may	516
suggest	285	might	75
indicate	150	could	55
suggesting	112	possible	32
ability	108	potential	23
indicated	99	possibility	10
appears	88	possibly	10
able	86	potentially	10
indicating	72	perhaps	5
likely	52	propose	4

Table 6. Most common *Certainty Level* cue expressions

4.3.1.3 Polarity

As can be seen in Table 7, only a small number of events are negated (6.1%). However, it is vital that such information is detected, as negation completely alters the meaning of the event.

Polarity	Frequency	% of total events
Positive (default)	34595	93.9%
Negative	2263	6.1%

Table 7. Distribution of annotated categories for *Polarity*

In the GENIA Event corpus, negation is an aspect of meta-knowledge that was annotated as part of the original annotation (via the *assertion* attribute). There is almost, but not complete agreement, between *Polarity=Negative* and *assertion=non-exist,* with a total of 2262 events annotated with the former and 2351 in the latter case. The slightly fewer negative annotations produced by our annotation are mainly due to the fact that some events annotated as negative in the original GENIA annotation actually convey low levels of interaction (rather than no interaction). An example is shown in sentence S84. As with previous example sentences, the event-trigger is underlined and the cue expression is emboldened.

S84 *AP-1 but not NF-IL-6 DNA binding activity was also detected in C5a-stimulated PBMC; however, its delayed expression (maximal at 4 hours) suggested a* ***less important*** *role in the rapid production of IL-8.*

The event encodes the fact that the expression of AP-1 only has a minor role (but not no role) in the rapid production of IL-8. As the GENIA annotation had no special means to encode that an event has low intensity or impact, the original annotator

chose to annotate this as a negative event, even though this is not strictly correct. Our annotation scheme, with its *Manner* dimension, allows the subtle difference between an event having a low impact or not happening at all to be encoded. Our scheme annotates low impact events such as the above as *Polarity=Positive* but *Manner=Low.*

In Table 8, we examine the distribution of negated events amongst the different *Knowledge Type* categories. Although negated events occur within events belonging to all *Knowledge Type* categories, the distribution is quite uneven. Only observations and analyses are negated with any amount of regularity. Events belonging to the remaining *Knowledge Type* values are virtually always expressed with positive polarity, with only around 3.5% of fact–bearing events being negative, and the other three categories (*Investigation, Method* and *Other*) only averaging one negative instance per hundred events.

Knowledge Type Category	**Negated events (% within category)**
Observation	1364 (10.6%)
Analysis	577 (8.7%)
Fact	105 (3.5%)
Other	187 (1.6%)
Method	10 (1.0%)
Investigation	20 (1.0%)

Table 8. Distribution of negated events among *Knowledge Type* categories

The low occurrence of negative instances amongst events with *Knowledge Type=Investigation* events is quite intuitive - it is the norm to investigate

why/whether something *does* take place, although in some instances there can be investigation into why something does not take place, such as in response to a previous negative finding, such as in S85.

S85 ***To determine*** *why alveolar macrophages* ***do not*** *express AP-1 DNA binding activity, ...*

Also for methods, it is highly unusual to say that a particular method was not applied, unless in contrast to the case where the method was applied, as the case in S86.

S86 *For comparison, we recruited a control group consisting of 32 healthy males and females with similar age distribution and* ***without*** *a history of exposure to MTBE or benzene.*

Table 9 displays the most commonly annotated cue expressions for negated events. Although the number of events we have identified as negated is roughly similar to those originally annotated in the GENIA Event corpus, our annotation has the advantage of having identified a suitable cue expression for each negated event.

Category	Frequency
not	1141
no	199
independent	113
without	65
failed	47
nor	47
absence	42

neither	38
unaffected	28
lack	23
un	23
unable	19
independently	18
resistant	15
fails	13

Table 9. Most common cue expressions for negative polarity

The word *not* constitutes around half of all cue expressions for negation (50.4%), and is over 5 times more common than the next most common cue expression, *no.* Although most of the words in the list have an inherently negative meaning, the third most common word, i.e. *independent* (together with its associated adverb *independently*), does not. Closer examination shows that this negative meaning is quite context-dependent, in that it only denotes a negative meaning for events of type *Correlation* and *Regulation* (together with its sub-type *Positive Regulation).* For *Regulation,* a typical example is shown in S87.

S87 *An alteration in the E2F-4 profile was **independent** of viral gene expression*

In S87, the word *independent* acts as both the event-trigger and the negative cue expression. The event denotes the fact that the alteration in the E2F-4 profile was not dependent on viral gene expression occurring. In other words, it is not the case that viral gene expression regulates the alteration in the E2F-4 profile. Events of type

Correlation are annotated when there is some kind of association that holds between entities and/or other events. Sentence S88 shows an example of both a positive *Correlation* event and a negated *Correlation* event.

S88 *LPS-induced NF-kappaB activation is protein tyrosine kinase dependent and protein kinase C* ***independent****.*

There are three relevant events in S88. Firstly, the word *induced* is the trigger for the *Positive Regulation* event in which *NF-kappaB activation* is regulated by *LPS.* The word *dependent* is the trigger for the second event, which shows that there is some kind of correlation between this positive regulation event and the protein *tyrosine kinase.* In contrast, the third event, triggered by *independent,* shows that no such correlation holds between the positive regulation and the protein *kinase C.* Hence, this is a negated *Correlation* event.

Some less commonly occurring negative cue expressions also only have negative meanings in very specific contexts. Consider S89:

S89 *These cells are deficient in FasL expression and apoptosis induced upon TCR triggering, although their cytokine (IL-2 and IFN-gamma) production is* ***normal****.*

In S89, the word *deficient* indicates a positive instance of a *Negative Regulation* event (i.e., the negative regulation does occur). However, the word *normal* indicates that no such negative regulation occurs in the case of IL-2 and IFN-gamma production. In the few instances where *normal* occurs as a negative polarity marker, it is used in similar contexts, i.e. to contrast with a previously stated *Negative Regulation*

event. The word *silent* appears to be usable in similar contexts to negate events of type *Positive Regulation*, in contrast to a positive occurrence of such an event.

4.3.1.4 Manner

As shown in Table 10, almost 5% of all events express a *Manner* value other than *Neutral*, which makes it only a slightly less commonly expressed phenomenon than negation. In the previous section, it has already been illustrated that the *Low* manner value can help distinguish between truly negative events, and those that occur at a low level or with low intensity. However, instances of *High* manner are much more common, and account for 81% of events for which there is an explicit indication of Manner.

Manner	Frequency	% of total events
Neutral (default)	35143	95.3%
High	1392	3.8%
Low	323	0.8%

Table 10. Distribution of annotated categories for *Manner*

The distribution of events annotated with either high or low Manner according to the *Knowledge Type* value of the event is shown in Table 11.

For the *Observation* category, explicit expression of *Manner* is observed in close to 1 in 10 events, making its frequency similar to the expressions of negation within this category. Of all events annotated for *Manner,* 66.5% correspond to those with the *Knowledge Type* value of *Observation*. This makes it clear that the main usage of *Manner* marking is to refine the descriptions of experimental observations and results.

Knowledge Type Category	**Events with *High* or *Low* Manner annotated (% within category)**
Observation	1141 (8.9%)
Analysis	276 (4.2%)
Fact	120 (4.0%)
Other	171 (1.5%)
Investigation	5 (0.2%)
Method	2 (0.2%)

Table 11. Distribution of negated events among *Knowledge Type* categories

Table 12 shows the 15 most common cue expressions for both the *High* and *Low* values of the *Manner* dimension. In both cases, most of the cue expressions consist of adjectives or adverbs, with a range of meanings referring to degree (e.g., *completely*), speed or rate (e.g., *rapidly*), strength or intensity (e.g., *strongly*) and level (e.g. *high*). These differences in meaning of the manner expressions can be explained by the varying semantics of the biological processes that are described by events. In most cases, items in the *High* manner list have counterparts in the *Low* list, e.g., *significant* vs. *little*, *high* vs. *low, strongly* vs. *weakly, completely* vs. *partially.* It is notable that a counterpart of *rapidly* (e.g., *slowly*) appears to be missing from the list of *Low* cue expressions.

In the *High* manner cue word list, a notable item is *overexpression.* Unlike the other cues in the list, which are independent of event type, this word is specific to events of type *Gene Expression,* as it combines the meaning of the event type with the expression of *High* manner. Comparable examples appear very rarely.

High Manner		Low Manner	
Cue	**Frequency**	**Cue**	**Frequency**
significantly	140	little	22
potent	84	low	15
markedly	81	little or no	13
rapidly	73	low levels	11
strongly	72	weak	11
rapid	65	limited	10
significant	39	low level	9
completely	36	weakly	9
strong	30	minimal	8
high	28	only a partial	8
high levels	28	no significant	8
over expression	26	partially	8
highly	23	barely	7
marked	23	to a lesser extent	6
dramatically	22	not significant	6

Table 12. Most common *Manner* cue expressions

Some of the annotated cues for both *High* and *Low* manner contain numerical values, meaning that a pattern matching approach may be required when trying to recognise them in unseen texts. For example, the expression *n-fold* is often used to denote *High* manner (often preceding the word *increase* or *decrease*), where *n* may be any numeric value. Otherwise, *by n%* may follow one of these words. To indicate *Low* manner, the expressions *n-fold less* or *n-fold lower* are sometimes used.

4.3.1.5 Knowledge Source

Regarding the *Knowledge Source* dimension, only 1.5% of events in total have any evidence that they come from a source other than the current study, as shown in Table 13. This low percentage may be expected, given that abstracts are meant to summarise the work carried out in the current study. In addition, citations, which are a common way to denote previous work, are often not allowed within abstracts. It should be noted that a considerably greater proportion of events marked as *Source=Other* would be expected when applying the scheme to full papers, in which the *Background* section will normally contain a large number of references to and descriptions of previous work (section 4.4.5). Of the events annotated as *Source=Other* within abstracts, the vast majority (86%) have the *Knowledge Type* value of *Analysis.*

Knowledge Source	Frequency	% of total events
Current (default)	36313	98.5%
Other	545	1.5%

Table 13. Distribution of annotated categories for Manner

Table 14 shows the 10 most commonly annotated cue expressions for *Source=Other.* Most of these consist of the words *previous* or *recent,* or phrases containing these words. The use of the passive voice with the present perfect tense (e.g. *has been studied)* is another common means to indicate that an event has previously been completed (e.g. in a previous study), but has relevance to the current study. This explains the relatively high occurrence of *has been* and *have been* as cues for *Source=Other*.

Cue	Frequency
previously	118
has been	89
recently	67
have been	39
previous studies	24
recent studies	17
recent	15
previous	14
our previous studies	10
earlier	6

Table 14. Most common cue expressions for *Source=Other*

4.3.1.6 Hyper-dimensions

Using the inference tables discussed earlier (section 3.3.6), we calculated the frequencies for the two hyper-dimensions, which are shown in Table 15.

Hyper-dimension	Category	Frequency	% of total events
New Knowledge	Yes	15985	43.4%
	No	20873	56.6%
Hypothesis	Yes	4924	13.4%
	No	31934	86.6%

Table 15. Distribution of categories for the two hyper-dimensions

As a comparison to these figures, the annotation carried out in [120] included annotating sentences containing descriptions of claims of new knowledge annotated in

chemistry and computational linguistics research articles. The results showed that the proportion of sentences containing new knowledge was 63% for the chemistry articles and 72% for the computational linguistics articles. It may be expected that the amount of new knowledge presented in biomedical research articles would be more similar to chemistry articles than computational linguistics ones. However, the proportion of *events* that represent new knowledge in our corpus is somewhat lower than the proportion of sentences that contain new knowledge in chemistry. This lower percentage can be explained in a number of ways. Firstly, unlike our scheme, [120] treat experimental methods as new knowledge, and these make up a significant proportion of the new knowledge in the chemistry articles. In any case, as has been reported above, abstracts have a different structure to articles, and experimental methods are rarely reported. In addition, our *New Knowledge* hyper-dimension takes certainty level into account, and excludes events which are highly speculative. However, certainty level is not taken into account in [120]. Finally, the granularity of the schemes is different. Whilst [120] annotates at sentence level, our annotation is at the event level, of which there are average of 3 to 4 per sentence. As some of these events represent non-propositional information, which cannot be treated as new knowledge, it makes sense that the proportion of events that represent new knowledge would be lower than the percentage of sentences that contain such information.

4.3.2 Inter-Annotator Agreement

In order to ensure the consistency and quality of the meta-knowledge annotation throughout the corpus, 104 randomly selected abstracts (10% of the entire corpus) were annotated by both annotators, allowing us to calculate their agreement rates.

For this purpose, the familiar measure of Cohen's kappa [121] was used, which adjusts the observed agreement for what would be expected by chance. If Cohen's kappa is represented by $\boldsymbol{k}$, the proportion of observed agreement by $\boldsymbol{p}$ and the proportion of expected agreement by $\boldsymbol{p_e}$, then the value of $\boldsymbol{k}$ can be calculated by the following formula:

$$k = (p - p_e) / (1 - p_e)$$

The results for each dimension are reported in Table 16.

Dimension	Kappa Value
Polarity	0.929
Source	0.878
Certainty Level	0.864
Manner	0.864
Knowledge Type	0.843

Table 16. Inter-annotator agreement rates

High levels of agreement were achieved in each dimension, with generally only very small differences between the agreement rates of different dimensions. This provides strong evidence that consistent annotation of meta-knowledge is a task that can be reliably undertaken by following the annotation guidelines.

The *Polarity* dimension has the highest rates of agreement. This could be because it is one of the two dimensions that have only two possible values (together with *Knowledge Source,* which has the second highest agreement rate). The two dimensions with three possible values (i.e. *Certainty Level* and *Manner*) have virtually identical rates of agreement, while *Knowledge Type* has the lowest agreement rate

(albeit only by a small amount). This is, however, to be expected – *Knowledge Type* has 6 possible values and in many cases, contextual information other than cue expressions is required to determine the correct value. Therefore, it can be a more demanding task than the assignment of other dimensions.

4.3.3 Annotation Discrepancies

We have studied the cases where there is a discrepancy between the two annotators. Whilst a number of these discrepancies are simple annotation errors, in which a particular dimension value was mistakenly selected during the annotation task, other discrepancies occur when a dimension value is identified by means of a cue expression that is not present in the list of sample cue expressions provided in the guidelines. In some cases, one of the annotators would notice the new cue, and use it to assign an appropriate category, but the other annotator would miss it. In order to minimise the occurrence of such cases, annotators were asked to flag new cue expressions, so that the lists of cue expressions in the guidelines could be updated to be as comprehensive as possible, and so ease the task of accurate annotation.

One of the largest areas of disagreement was between the *Knowledge Type* categories of *Observation* and *Fact.* For a number of reasons, distinguishing between these types can often be quite tricky, and sometimes there is no clear evidence to suggest which of the categories should be chosen. Events of both types can occur in the present tense, and explicit cue expressions are more frequently absent than present. Often, the extended context of the event (including possibly other sentences) has to be considered before a decision can be made. In some cases, it appears that domain knowledge is required to make the correct decision.

In the remainder of this section, we look at some particular cases of annotation discrepancies, some of which appear to be influenced by the expertise of the annotator.

Long sentences seemed to prove more problematic for the biologist annotator, and meta-knowledge information was sometimes missed when there is a large gap between the cue expression and the event-trigger. Consider sentence S90 (below), in which the word *indicated* should cause *both* the event with the trigger *prevented* and the one with the trigger *activated* to be annotated with *Knowledge Type=Analysis*.

S90 *Accordingly, electrophoretic mobility shift assays (EMSAs)* ***indicated*** *that pyrrolidine DTC (PDTC) prevented NF-kappaB, and NFAT DNA-binding activity in T cells stimulated with either phorbol myristate acetate plus ionophore or antibodies against the CD3-T-cell receptor complex and simultaneously activated the binding of AP-1.*

Whilst it is straightforward to understand that *indicated* affects the interpretation of the event-triggered by *prevented*, it is less easy to spot the fact that it also applies to the event triggered by *activated,* due to the long description of the T cells, which precedes this trigger.

It appears that having some linguistic expertise is an advantage in order to cope with such cases. The biologist would often fail to consider a cue word as potentially affecting the interpretation of an event unless it occurred in close proximity to the event itself. In contrast, the linguist would normally detect long distance dependencies between cue expressions and triggers without difficulty. This is to be expected, given that the linguist is familiar with grammatical rules. However, given the generally high levels of agreement, such complex cases appear to be reasonably rare.

Other annotation discrepancies reveal further differences in the approaches of the annotators. Whilst some grammatical knowledge appears to be advantageous, using a purely grammatical approach to the recognition of meta-knowledge is not always correct. The semantic viewpoint appears to be the one most naturally taken by the biologist annotator, as is evident in sentences such as S91:

S91 *This study **demonstrates** that GC act as a primary inducer of sialoadhesin expression on rat macrophages, and that the response can be enhanced by IFN-beta, T cell-derived cytokines, or LPS.*

In S91, we focus on the events triggered by *inducer* and *enhanced,* which are of type *Positive Regulation.* The word *demonstrates* is a cue expression for the *Knowledge Type* category *Analysis.* Taking a purely grammatical approach, the word *demonstrates* affects the interpretation of the verbs *act* and *enhanced.* Accordingly, both annotators marked the event triggered by *enhanced* as *Knowledge Type=Analysis*. However, the biologist also annotated the *inducer* event with *Knowledge Type=Analysis,* also marking *demonstrates* as the cue expression. Considering semantics, this is correct – the actual meaning of the first part of the sentence is that *This study demonstrates that GC induces sialoadhesin expression on rat macrophages*.

Sentence S92 illustrates the need to carefully consider the meaning of words and phrases in the context of the event, as well as simply looking for relevant keywords.

S92 *Changes of any cysteine residue of the hRAR alpha-LBD had **no significant** influence on the binding of all-trans RA or 9-cis RA.*

One of the annotators had annotated the *Regulation* event with the trigger *influence* with *Polarity=Negative* (cue word: *no)* and *Manner=High* (cue word: *significant*). However, this is incorrect - it is the word *significant* that is negated, rather than the event itself. As *significant* would normally be a marker of *High* manner, negating it means that it should be treated as a *Low* manner marker. Accordingly, the other annotator correctly identified *no significant* as the cue phrase for *Manner=Low,* with the polarity of the event correctly remaining positive.

The interplay between events in the GENIA event corpus can be complex, especially as events can occur that have no trigger phrase. The links between different events in a sentence often have to be understood before a decision can be made about which of the events a particular piece of meta-knowledge should apply to. In such cases, a detailed understanding of the domain could be considered to be an advantage. The following sentence fragment (S93) illustrates such a case, in which *absence* constitutes a cue expression for *Polarity=Negative* for one of the events.

S93 *In the **absence** of TCR-mediated activation, Vpr induces apoptosis...*

Three events have been identified as part of the original GENIA Event annotation:

1. A *Positive Regulation* event with the trigger *mediated* (i.e., positive regulation of activation by TCR). At first glance, it is to this event that the negative polarity appears to apply.
2. A second *Positive Regulation* event, with the trigger *induces* (i.e. positive regulation of apoptosis by Vpr)
3. A *Correlation* event with no trigger, providing a link between the first two events (1 and 2 above). In fact, the negative polarity applies to this event.

> The event conveys the fact that Vpr induces apoptosis even when there is no TRC-mediated activation, indicating that there is no correlation between these two events.

The above examples demonstrate that accurate meta-knowledge annotation can be a complex task, which, according to the event in question, may have to take into account the structure and semantics of the sentence in which the event is contained, as well as the semantics of the event itself and possibly the interplay between events.

Our inter-annotator agreement results suggest, however, that the task of meta-knowledge annotation can be accurately undertaken, given appropriate guidelines and training. Furthermore, the results provide evidence that high quality meta-knowledge annotations can be produced regardless of the expertise of the annotator. Although we have highlighted certain cases where either domain knowledge or linguistic expertise appears to be a distinct advantage, neither seems to be a prerequisite. This is in agreement with [116], in which biologist annotators were trained to carry out linguistically-motivated annotation of biomedical events, with good levels of agreement.

4.4 Annotation of Full Papers

In order to investigate the scalability of our meta-knowledge scheme, we conducted a case study to investigate the feasibility of applying it to full papers. Although the design of our scheme was originally guided only by reference to abstracts, such scalability is important given that work on event extraction is gradually being scaled from abstracts to full papers, and also that the automatic recognition of meta-knowledge about events can be highly useful for building more sophisticated infor-

mation extraction systems. Our case study involved the annotation of 4 full papers using the same meta-knowledge annotation guidelines that were used to create the GENIA-MK corpus. We refer to the resulting (meta-knowledge enriched) corpus as FP-MK corpus. These full papers had already been annotated with bio-event information using the GENIA Event annotation guidelines. The annotations were performed by a single annotator with a strong computational linguistics background, who had previously been involved in the design and implementation of the meta-knowledge annotation scheme. No specific difficulties were encountered in applying the scheme to events in full papers. Furthermore, the results strongly suggest that the existing meta-knowledge annotation scheme can be successfully applied to full papers, without any modifications.

Table 17 summarises the distribution of the annotations in the FP-MK corpus amongst the different categories for each dimension, and Table 18 shows the most frequent cues for each category together with their relative frequencies, i.e., the percentage of events of the specified category in which the cue is annotated. In the remainder of this section, we provide a brief discussion of these annotation results.

Dimension	Category	Frequency	Relative Frequency
Knowledge Type	Analysis	381	22.3%
	Investigation	65	3.8%
	Observation	619	36.2%
	Fact	70	4.1%
	Method	100	5.8%
	Other	475	27.8%
Certainty Level	L1	39	2.3%
	L2	162	9.5%
	L3	1509	88.2%
Polarity	Negative	63	3.7%
	Positive	1647	96.3%
Manner	High	66	3.9%
	Low	15	0.9%
	Neutral	1629	95.3%
Source	Current	1369	80.1%
	Other	341	19.9%
Hyper-Dimensions	New Knowledge	489	28.6%
	Hypothesis	259	15.1%

Table 17. Category distributions for all dimensions

Dimension	Category	Most Frequent Cues and their RF
Knowledge Type	Analysis	show (16%), demonstrate (14%), indicate (9%), suggest (7%), reveal (5%), can (4%), thus (3%), may (3%)
	Investigation	determine (19%), analyze (15%), elucidate (11%), evaluate (9%), detect (5%), indicate (5%), test (5%), examine (3%), investigate (3%)
	Observation	observe (4%), find (3%), show (1%), document (1%), exhibit (1%)
	Fact	known (6%), well established (3%), well known (2%), fact (2%)
Certainty Level	L1	may (54%), can (15%), possibility (10%), not clear (5%), not understood (5%)
	L2	indicate (22%), can (15%), suggest (11%), ability (6%), able (6%), potential (4%), hypothesize (3%), imply (3%), suspect (3%)
Polarity	Negative	not (57%), no (18%), failure (10%), non (8%), fail (2%), inability (2%)
Manner	High	significantly (17%), well (12%), much (11%), n-fold (9%), strong (9%), strongly (6%), high (3%), higher (3%)
	Low	minimal (13%), little (13%), weak (13%), weaker (13%), n% (7%), less (7%)
Knowledge Source	Other	Citation (78%), has been (12%), previously (2%), recently (2%)

Table 18. Most frequent cues for each category

4.4.1 Knowledge Type

The most common annotated category is *Observation*, constituting just over a third of the total number of events. This is unsurprising, since a large part of most biomedical papers would be expected to report on definite experimental observations and results, both of which fall into this category.

Considering individual sections within the full papers, *Observation* events are most prevalent in *Background* (42% of all events in this section category). It may seem slightly surprising that the frequency of *Observation* events in *Background* is a greater than in *Results* sections. However, *Observation* events can refer to previous work as well as current work, and *Background* sections often refer to findings from a large number of related studies. In the *Results* sections, approximately 36% of events describe observations; while in the *Discussion* section, the frequency of such events is even lower (32%). This is to be expected, since greater proportion of this section type would normally be analytical in nature.

Only in a small fraction (12%) of *the Observation* events is the *Knowledge Type* value determined by the presence of an explicit lexical cue (mostly sensory verbs). In most cases, the tense of the event-trigger and the context of the event (both local and global position within the paper) were found to be important factors.

The second most prevalent category is *Other*. These events generally constitute participants of other events whose *Knowledge Type* value is either *Investigation*, *Analysis* or *Fact*. Out of the context of their parent event, these participant events have no specific *Knowledge Type* interpretation. No explicit lexical cues were annotated for this category.

A relatively large proportion of events (more than one fifth) belong to the *Analysis* category. This makes sense, given that analytical elements are normally to be found to some extent in most section types in full papers. These include the *Background* section, where such events are most likely to provide overviews or interpretations of previous work, as well the *Results*, *Discussion* and *Conclusions* sections, where analyses, interpretations and conclusions regarding authors' own work most commonly appear. As may be expected, the frequency of *Analysis* events is highest in *Discussion/Conclusion* sections, where they constitute over one quarter (27%) of all events.

An explicit lexical cue was found for each *Analysis* event. The cues comprised verbs, modal auxiliaries and certain adverbs (such as, *thus* and *therefore*).

Almost 6% of the events belong to the *Method* category. Although full papers generally include a fairly large *Methods* section, the small number of events falling into this category is largely a consequence of the fact that the GENIA Event annotation focusses on dynamic relations, i.e., cases where at least one of the biological entities in the relationship is affected, with respect to its properties or its location, in the reported context. This means that descriptions of methods are often less relevant in the GENIA Event annotation than are events describing observations and analyses.

Our case study suggests that only a small proportion of events in full papers (around 4%) describe factual knowledge. Such events are not evenly distributed throughout papers, and occur most frequently in the *Background* section (7.5% of all events in this section type), in order to provide context for the new research described in the paper. They can also appear in the *Discussion* sections (4.5% of events), where they may be contrasted or compared with the outcomes of the current study. As may be

expected, factual knowledge is almost never referred to in the *Results* sections of papers. Similarly to the *Observation* category, most (85%) events from this category do not have an explicit lexical cue.

The *Investigation* category is the least frequent *Knowledge Type* in full papers. The results of our annotation experiment suggest that *Background* sections will normally very briefly introduce the subject of investigation (2.5% of events in this section type). A slightly more detailed description of the investigation is then given in *Results* sections (5.4% of all events in this section type). It is also possible that the aim of the research will be very briefly reintroduced in the *Discussion* section of the paper (an average of 1.8% of all events in this section type). All *Investigation* events are accompanied by an explicit lexical cue.

4.4.2 Certainty Level

Almost 12% of all events in our full paper sample are expressed with some degree of uncertainty. All uncertain events belong to the *Knowledge Type* category *Analysis*. Furthermore, 43% of all *Analysis* events are annotated as having slight speculation (*Certainty Level = L2*), whilst 10% are reported with a larger degree of speculation (*Certainty Level = L1*). The marking of uncertainty is sometimes necessary in scientific research literature. The author's analyses of experimental results may have produced important outcomes, but yet the authors are not confident that their analysis is completely reliable. As previously mentioned (section 4.3.1.2), it has been shown [95] that authors tend to avoid higher levels of speculation (*Certainty Level = L1*) as this would reduce the credibility of their analyses. However, they insure against overstatement by using slight hedging (*Certainty Level = L2*).

Considering individual sections helps to confirm Hyland's statement. Although the proportion of *Analysis* events that is assigned a *Certainty Level* value of *L1* is fairly constant across the *Background, Results* and *Discussion* sections, the proportions of *L2* events have more variation. The relative frequency of such events is lowest in the *Background* sections (36% of *Analysis* events). Since this section deals mainly with reporting the work of others, there is perhaps less need to hedge, as it is not the authors' own credibility at stake. In contrast, the relative frequency of slightly hedged *Analysis* events is noticeably higher in the *Results* and *Discussion* sections (46% and 51%), respectively, where the authors' own work is the main focus, and hence interpretations and analyses of results are often stated more tentatively.

In terms of cues for events with non-default *Certainty Level* values, modal auxiliaries account for most (70%) of the *L1* events, while the cues for *L2* include both verbs and modals.

4.4.3 Polarity

Just under 4% of all events in the FP-MK corpus are negated. Almost all of these events belong to the *Knowledge Type* categories of *Observation* or *Analysis*, which is fairly intuitive. One would not, for example, expect to encounter many cases where *Investigation* or *Method* events are negated, . The distributions of negated events vary across different sections of the full papers. The proportions encountered in *Background* and *Discussion* sections are quite similar to each other (around 2% in each section), compared to around 6% of negated events in *Results* sections. Thus, it appears that it is very rare for anything other than positive results to be mentioned in the former 2 section types. In contrast, when reporting directly on one's own experimental results, negative results are mentioned more frequently.

Although several negation cues were annotated, the adverbial *not* accounts for over half of the negated events.

4.4.4 Manner

Almost 5% of all events in the full-paper sample are expressed with a *Manner* other than *Neutral*. This proportion is fairly constant throughout the *Background, Results* and *Discussion* sections of the full papers, showing that, although fairly rare, information about the manner of events can be of relevance to the discussion in different parts of the paper. However, the expression of *High* manner is 4 times more frequent than that of *Low* manner. Similarly to negation, most *Manner=High* events belong to *Knowledge Type* categories of *Observation* or *Analysis*.

Another similar pattern to the *Polarity* dimension is that instances of events with a *Manner* value of *Low* seem to appear with any regularity only in the *Results* sections of the papers, where they appear with just over half the frequency of events whose *Manner* value is *High.* In contrast, the *Low* value was never annotated in the *Background* sections of the papers, and was only annotated for less than 1% of events in the *Discussion* sections. This suggests that authors might ascribe more importance to *High* manner events, and may consider *Low* manner events to be less significant. This hypothesis is further strengthened by the fact that there is a degree of similarity between the *Low* manner events and the negated events, and historically, negated results have been considered less important [122]. However, this trend has been changing recently (see section 5.1).

Most manner cues are adverbs or adjectives; however, similarly to abstracts, numerical values (such as, *n-fold* and *n%*) are also used to express *High* manner.

4.4.5 Knowledge Source

Nearly 20% of all events in the full papers belong to the *Other* category. The concentration of such events is highest in the *Background* sections of the papers, where over 40% of the events are attributed to other sources. This is expected, since it is normally in the *Background* section where one encounters the highest concentration of descriptions of previous work. The *Discussion* sections of the papers also have a high (over 25%) concentration of *Other* events, since in this section, it is common to compare and contrast the outcomes of the current work with those of previous, related studies. The frequency of *Other* events in the remaining sections is considerably lower. For example, in the *Results* sections of the papers considered, less than 7% of events are annotated as *Other*. While citations accounted for most of the *Other* events, the use of past perfect tense and explicit markers (such as *previously* and *recently*) also serve as cues.

4.4.6 Hyper-dimensions

Using the annotations for *Knowledge Type*, *Certainty Level* and *Source* dimensions, we computed the values for the *New Knowledge* and *Hypothesis* dimensions. We found that nearly 29% of all events conveyed new knowledge, and over 15% of all events represented hypotheses. Events conveying new knowledge were predominantly found in the *Results*, *Discussion* and *Conclusion* sections, while hypotheses were also found in these sections, as well as in the *Background* section. The *Methods* section contained hardly any hypotheses or claims of new knowledge.

4.5 Comparison of Abstracts and Full Papers

In this section, we compare the distribution of meta-knowledge annotations obtained

from our case study of full papers with those obtained for abstracts. Table 19 shows the difference between the category distributions for full papers and abstracts. A brief discussion of the differences in each dimension is as follows:

Dimension	**Category**	**Difference in Relative Frequencies in Full Papers (FP) and Abstracts (A): RF(FP) – RF(A)**	**% Change in Relative Frequency: |RF(FP) – RF(A)| / min(RF(FP), RF(A))**
Knowledge Type	Analysis	4.4%	24.8%
	Investigation	-1.5%	39.0%
	Observation	1.4%	4.1%
	Fact	-4.0%	98.7%
	Method	3.2%	120.8%
	Other	-3.5%	12.7%
Certainty Level	L1	0.2%	9.7%
	L2	3.5%	57.6%
	L3	-3.7%	4.2%
Polarity	Negative	-2.5%	66.7%
	Positive	2.5%	2.6%
Manner	High	0.1%	2.2%
	Low	0.0%	0.0%
	Neutral	-0.1%	0.1%
Knowledge Source	Current	-18.5%	23.1%
	Other	18.5%	1248.6%
Hyper-Dimensions	New Knowledge	-14.8%	51.7%
	Hypothesis	1.8%	13.4%

Table 19. Difference between the category distributions for full papers and abstracts

4.5.1 Knowledge Type

The biggest difference is seen for the *Method* events, which are more than twice as abundant (in terms of relative frequency) in full papers as in abstracts. This is probably because abstracts tend to focus more on the results and their significance, rather than how these results were obtained. However, owing to the previously explained "dynamic" nature of GENIA Events, the frequency of *Method* events is quite low even for full papers.

A further feature of abstracts is that they tend to contain one or two sentences summarising current knowledge (i.e., well known facts) in the relevant field. Since the average size of abstracts in the GENIA Event corpus is 9 to 10 sentences [29], the relative frequency of facts in abstracts is quite high (over 8%). This proportion is comparable to the number of factual events in *Background* section of full papers (over 7% of all events in this section type), where the current state of knowledge is also discussed in some detail. However, events describing facts are far scarcer in the other sections of full papers and, given their overall length, the relative frequency of *Fact* events in full papers as a whole is only around half of the frequency in abstracts.

Regarding *Investigation* events, their relative frequency in the *Results* sections of the full papers is comparable to their relative frequency in abstracts (around 5%). However, in the same way as the *Fact* category, the extremely rare appearance of *Investigation* events in other sections of full papers means that overall relative frequency in full papers is again much lower than in abstracts.

The relative frequency of *Analysis* events is around 25% higher in full papers than in abstracts. In contrast to *Fact* and *Investigation* events, *Analysis* events are found

with quite high frequency in several sections of full papers. For the *Other* and particularly the *Observation* categories, there is much less variation between the relative frequencies in full papers and abstracts. Thus, clear reporting of experimental observations is equally important throughout both full papers and abstracts,

4.5.2 Certainty Level

Owing to the very nature of abstracts, a high proportion of events with no uncertainty is to be expected. As explained in section 4.2, authors aim to "sell" the most positive aspects of their work in abstracts. Therefore, it makes sense that the majority of analyses should be presented in a confident manner. However, authors tend to be more cautious while detailing their results and findings in the main body of papers, in order to maintain credibility in case their results are later disproved. The fact that the proportion of slightly hedged *Analysis* events is particularly high in the *Results*, *Discussion* and *Conclusion* sections of full papers (rising as high as 51% in the *Discussion* sections) helps to explain why *L2* events are more than 50% more frequent in full papers than in abstracts. The relative frequency of *L1* events is also higher in full papers by about 10%.

4.5.3 Polarity

Interestingly, the relative frequency of negated events is significantly (67%) higher in abstracts than in full papers. This can partly be explained by the fact that negative results are sometimes more significant than positive results [122], and are, therefore, highlighted in the abstracts. In addition, since negated events only appear with any regularity in the *Results* sections of full papers, this helps to explain their lower relative frequency than in abstracts when the complete paper is considered.

4.5.4 Manner

The distribution of *High* and *Neutral* manner is very similar in abstracts and full papers, and the distribution of *Low* manner is exactly same. This follows the same trend described in section 4.4, where it was also noted that the proportions of events with explicit manner markings are also fairly similar across several individual section types within full papers.

4.5.5 Knowledge Source

This is the dimension for which the largest difference in category distribution exists between abstracts and full papers. Full papers contain 12.5 times as many *Other* events as abstracts. This is mainly because abstracts are meant to summarise the work carried out in the current study. Furthermore, citations, which are the most common way to denote previous work, are often not allowed within abstracts. In contrast, full papers normally mentioned related work quite extensively, most notably in *Background* and *Discussion* sections.

4.5.6 Hyper-Dimensions

While the relative frequency of *Hypothesis* events is higher in full papers, the proportion of *New Knowledge* events is significantly higher in abstracts. This is mainly because, in abstracts, authors typically include most of new discoveries and results, while only mentioning the main hypotheses.

4.6 Conclusion

We designed our meta-knowledge annotation scheme to enrich corpora of biomedical events with information about their characterisation or interpretation, based on

their textual context. The scheme was designed to be portable, in order to allow integration with the various different schemes for event annotation that are currently in existence. In this chapter we have described the application of the meta-knowledge annotation scheme to two corpora of bio-events. As a first major annotation effort, the scheme was applied to the largest currently available corpus of biomedical events (i.e. the GENIA Event corpus) to create the meta-knowledge enriched GENIA-MK corpus. Inter-annotator agreement rates of between 0.84-0.93 Kappa (according to annotation dimension) show that high levels of annotation quality and consistency can be achieved by following the annotation guidelines. Furthermore, it appears that, subject to the provision of these guidelines and a suitable training programme, meta-knowledge annotation can be performed to a high standard by annotators without specific areas of expertise, as long as they have a good command of the English language.

Further to the creation of the GENIA-MK corpus, we conducted a case study to investigate the feasibility of applying the annotation scheme to full papers. This is important, given that work on event extraction is gradually being scaled from abstracts to full papers. Our case study involved the creation of the FP-MK corpus through meta-knowledge enrichment of bio-events in 4 full papers, which had already been annotated with bio-event information using the GENIA event annotation guidelines. The results of the case study strongly suggest that the existing meta-knowledge annotation scheme can be successfully applied to full papers, without any modifications.

Chapter 5: Polarity of Bio-events

In this chapter, we provide details of the first comprehensive study on the analysis and identification of negated bio-events. We begin with an introduction to the task of identifying negated bio-events. We present a typology of negated bio-events, which has been derived from a detailed analysis of the three open access bio-event corpora containing negation information i.e., GENIA Event, BioInfer and BioNLP'09 ST. We then analyse the key aspects of a machine learning solution to the problem. These include the selection of negation cues, feature engineering and the choice of learning algorithm. Our analysis has been informed by a series of experiments involving four different lists of negation cues, four main sets of features and six learning algorithms. We used 10-fold cross validation for all experiments. Combining the best solutions for each aspect of the problem, we propose a novel framework for the identification of negated bio-events. We have evaluated our system on all three open access corpora of negated bio-events. It performs consistently on all corpora. It significantly surpasses the previously reported best results on the BioNLP'09 ST corpus, and achieves even better results on the GENIA Event and BioInfer corpora, both of which contain more varied and complex events.

5.1 Introduction

Negation is considered a universal property of all human languages [123]. However, the concept and manifestation of negation in natural languages is far more subtle and complex in force and scope than it is in formal logic [124-126]. Nonetheless, negation occurs frequently in scientific literature, especially in the domain of biomedi-

cine. Vincze et al. [107] report that around 13% of sentences found in biomedical research articles are negated.

Historically, in the field of biomedical text mining, the main motivation for the identification of negated events had been to ensure their exclusion from extracted lists of interactions. This was mainly because most biomedical research has been focused around the publication and analysis of positive results [122]. However, recently, there has been a growing interest in negative results, for example:

- The Journal of Negative Results in Biomedicine [127] has been launched, which, as the name suggests, focuses specifically on negative results.
- The Negatome database [128] has been released, which provides information about non-interacting protein pairs.
- Efforts have been made to incorporate negation into popular biomedical ontologies [129].

Recently, negation detection has been identified as the foremost challenge in biomedical relation extraction [130]. More specifically, it has been argued that the recognition of negated bio-events is of fundamental practical significance for researchers in most biomedical disciplines [131].

5.1.1 Negated Bio-events

Vincze et al [107] define negation in the context of biomedical literature as “the implication of the nonexistence of something”. Negated events have been identified in some bio-event corpora; although an explicit definition of a negated event has not been supplied, the implicit definition equates negation with non-existence. The indi-

cation of non-existence could be explicit (e.g., the presence of a negation marker) or implicit (e.g., semantic inference).

Of the bio-event corpora mentioned in section 2.1.3, only three contain information about event polarity; these are GENIA Event, BioInfer and BioNLP'09 ST. Negation cues have been explicitly identified only in BioInfer. Table 20 shows the relevant statistics for the three corpora. In terms of volume, the GENIA Event corpus is the largest, with almost 37,000 events, while BioInfer is the smallest with fewer than 2,700 bio-events. In terms of event-types, BioInfer is the richest, with 60 event-types and BioNLP'09 ST is the simplest, with only 9 event-types. Interestingly, the distribution of negated bio-events in all three corpora is fairly uniform, ranging between 6.1% and 6.4%.

Corpus	Event Types	Total Events	Number of Negated Events	Percentage of Negated Events
GENIA Event	36	36,858	2,351	6.4%
BioInfer	60	2,662	163	6.1%
BioNLP'09 ST	9	11,480	722	6.3%

Table 20. Statistics for bio-event corpora containing polarity information

5.1.2 Identification of Negated Bio-events: Task Description and Analysis

Following previous work [132-136], we treated the task of identifying negated bio-events as an independent task in itself. That is, we assumed that the event annotation has already been performed, and aim to find automated means of classifying these events according to their polarity.

A related negation detection task, which has received significant attention recently, is the detection of negation scopes [137]. This involves the identification of the sequence of words in a sentence which is affected by a negation cue. Despite the apparent similarities, identification of negated bio-events is essentially different from negation scope detection. While scope annotation focuses on linguistic properties of the text, the goal of bio-event annotation is to identify which kinds of biological information appear in which parts of the text and how they are related. Therefore, the expression of bio-events in text has two distinguishing characteristics [29, 31, 49]:

1. Bio-event annotation is information-centred and depends entirely on the biologists' conception of the relationship between an event, its participants and other events expressed in the text.
2. The event-trigger and participants of an event are each mapped to a different span of text. This means the description of an event is usually spread over several discontinuous spans in text, which could belong to different clauses within a sentence.

In contrast to the above characteristics, the scopes of negation cues are continuous and relatively less ambiguous [107]. A few interesting consequences of this contrast are:

- A sentence containing a negation cue may not contain any negated events at all.
- At the other extreme, certain events may be negated even when a negation cue is not present in the sentence. This point is discussed further in section 5.3.1.

- The event-triggers and/or the participants for many events may fall under the scope of a negation cue; however, it is highly unlikely that all of these events will be negated.

Vincze et al. [138] conducted an in-depth comparison of a linguistically annotated corpus of negation scopes (BioScope) and a biologically annotated corpus of negated bio-events (GENIA Event). They found that only half (51%) of the bio-events with event-triggers inside the scope of a negation cue were actually negated. Conversely, 16% of the negated bio-events had event-triggers which were outside the scope of the negation cues present in the sentence. They concluded that negation scope detection is not sufficient for the identification of negated bio-events, as the latter is a more complex task.

Based on the above discussion, we conclude that the identification of negated bio-events requires a deeper and more complex analysis than other negation detection tasks like negated term detection, negated PPI detection and negation scope detection.

5.2 Related Work

Negation detection has been a neglected area in open-domain natural language processing, and most research has been performed in the biomedical domain [139]. This section provides a brief overview of the previous work done on types of negation, negation cues, detection of negated terms and negation scopes, detection of negated PPIs and identification of negated bio-events.

5.2.1 Types of Negation

One of the first attempts at classifying negation in natural language was made by Aristotle. He concluded that negations can be divided into four types, which he named as correlation (e.g., *double* vs. *half*), contrariety (e.g., *good* vs. *bad*), privation (e.g., *blind* vs. *sighted*) and contradiction (e.g., *he sits* vs. *he does not sit*) [125]. In terms of more recent work, Tottie [124] presented a taxonomy of clausal negations in English. She identified 6 top-level categories of clausal negation as: denials, rejections, imperatives, questions, supports and repetitions. Harabagiu et al. [140] identified two main classes of negation: directly licensed negations and indirectly licensed negations. The directly licensed negations include: overt negative markers (such as *not*), negative quantifiers (like *no*) and strong negative adverbs (like *never*). The indirectly licensed negations include: verbs or phrasal verbs (such as *fail*), prepositions (such as *without*), weak quantifiers (such as *few*) and traditional negative polarity items (such as *a red cent*). Huang and Lowe [141] proposed a classification of negations found in medical reports. Their classification was based on the syntactic category of the negation signal and phrase patterns. They identified 4 syntactic categories of negation signals: adjective-like (such as *no*, *absent* and *without*), adverb (such as *not*), verb (such as *deny*) and noun (such as *absence*). They also identified 9 phrase patterns corresponding to the syntactic categories.

Sanchez-Graillet and Poesio [113] analysed negated PPIs in 50 biomedical articles. They identified seven classes of negation for PPIs. This classification is based on lexical and syntactic patterns; however, it is specific for PPIs and cannot be trivially extended to all types of bio-events.

5.2.2 Negation Cues

Chapman et al. [142] compiled a comprehensive list of 272 negation cues specific to medical discharge summaries. Mutalik et al. [126], despite identifying over 60 cues, report that only a small set of negation cues account for most of the negation instances. In their corpus of 40 medical documents, only four negation cues accounted for almost 93% of all negation instances. These cues are *no* (49%), *denies/denied* (21%), *not* (13%) and *without* (10%). Similarly, Tolentino et al. [143] analysed negated biomedical concepts occurring in a corpus of 41 medical documents. They found that only 5 negation cues (*no*, *neither/nor*, *ruled out*, *denies* and *without*) account for 89% of all negated concepts found in the corpus. Elkin et al. [144] created an ontology of terms that start negation (e.g., *no*, *denies* and *ruled out*) and another set which stop the propagation of the assignment of negation (e.g., *other than*). Kilicoglu and Bergler [132] created a list of 9 negation cues from the BioNLP'09 ST corpus. Morante [145] compiled a list of negation cues observed in the BioScope [146] corpus, identifying 8 ambiguous and 21 unambiguous negation cues. She also provided a description for the scope of each cue based on its syntactic context. Sarafraz and Nenadic [136] used previous studies on negation to derive a primary list of 14 negation cues. They further compiled a secondary list of 18 additional negation cues that were semi-automatically extracted from the BioNLP'09 ST corpus. Interestingly, their list contains the word *inhibit*, which is treated as an indicator of *negative_regulation* (and not negation) in the BioNLP'09 ST, GENIA Event and BioInfer corpora.

In terms of automated approaches, Morante and Daelemans [139] proposed a machine learning system for the identification of negation cues. Their system achieved

an F-score of over 99% for both clinical notes and biomedical abstracts. However, their system treated 17 strings as unambiguous negation markers i.e., every occurrence of these strings was treated as a negation cue. These unambiguous cues accounted for 95% of all instances of negations. Agarwal and Yu [147] developed a system for the automatic identification of negation cues using Conditional Random Fields (CRF). Their system achieved an F-score of 98% for clinical notes and 97% for biomedical abstracts.

5.2.3 Detection of Negated Terms and Negation Scopes

The bulk of work on negation detection in the biomedical domain has been focused on the detection of negated terms in medical reports. This includes both rule-based and machine learning approaches. The key rule-based solutions include those presented by Chapman et al. [142], Mutalik et al. [126], Elkin et al. [144], Huang and Lowe [141] and Boytcheva et al. [148]. The key machine learning approaches include the systems presented by Averbuch et al. [149], Goldin and Chapman [150], Goryachev et al. [151], Rokach et al. [152] and Councill et al. [153].

Vincze et al. [146] developed BioScope, an open access corpus of biomedical text containing token level annotations for negation cues and their respective scopes. The BioScope corpus comprises three sub-corpora: (1) clinical reports containing 6,383 sentences, (2) biomedical articles containing 2,670 sentences, (3) biomedical abstracts containing 11,871 sentences. Morante and Daelemans [139] presented a machine learning approach for detecting the scope of negation cues, and tested their system on the BioScope corpus. Their system determined the full scope of negation cues with an accuracy of 66% for abstracts, 41% for papers and 71% for clinical notes.

5.2.4 Detection of Negated PPIs

Sanchez-Graillet and Poesio [113] developed a set of heuristics for extracting negated PPIs from biomedical articles. They implemented their system using a Functional Dependency Grammar (FDG) parser. Their preliminary results range from 54% to 63% F-score, depending on the method of protein name recognition. The system achieved 77% F-score when used with gold standard protein annotations.

5.2.5 Detection of Negated Bio-events

Identification of negated bio-events was an optional sub-task in the BioNLP'09 Shared Task Challenge [154]. Six teams participated in this task and reported the first results on the identification of negated bio-events. Kilicoglu and Bergler [132] achieved the best results with their rule-based system. They achieved 14% recall, 51% precision and 23% F-score. Van Landeghem et al. [135] obtained the second best results with 11% recall, 45% precision and 17% F-score. They also used a customised rule-based system. MacKinlay et al. [134] used a machine-learning approach with complex deep parse features. Their system achieved the third best results with 5% recall, 34% precision and 9% F-score. It is important to note that these systems did not use gold standard event annotations as input. Instead, they performed both event extraction and identification of negated events. The approximated F-scores for these systems if they were to detect negations on gold standard event annotations are 38%, 26% and 28%, respectively. These values have been calculated using a linear extrapolation function and the maximum (100%) recall value for event extraction.

Sarafraz and Nenadic [136] proposed a machine learning approach for the identification of negated bio-events. They implemented an SVM classifier with a linear kernel

using features engineered from a sentence parse tree with lexical cues. They trained their classifier on the BioNLP'09 Training dataset and tested on the BioNLP'09 Development dataset. They achieved 38% precision, 76% recall and 51% F-score. In a further experiment, they split the data into smaller datasets according to event-types, and trained and tested the classifier separately for each smaller dataset. This way, they achieved a micro average of 49% precision, 88% recall and 63% F-score.

5.3 A Typology of Negated Bio-Events

The analysis presented in section 5.1.2 mandated further investigation into the causes and types of negation in bio-events. We conducted an in-depth analysis of the manifestations of negation observed in the three open access bio-event corpora containing negation information. We analysed a total of 1,000 randomly selected negated bio-events; of which, 600 negated bio-events were from the GENIA Event corpus (over 25% of all negated events in the corpus), 300 negated bio-events were from the BioNLP'09 Shared Task corpus (over 40% of all negated bio-events in the corpus) and 100 negated bio-events were from the BioInfer corpus (over 60% of all the negated bio-events in the corpus).

Our analysis revealed several causes and types of negation, which we have grouped together to formulate a typology of negated bio-events based on the relationships between negation cues and individual event constituents. Our typology consists of five classes, of which the first four classes are always expressed in the text through the use of an explicit negation cue, whereas the manifestations of the final class lack explicit negation cues.

5.3.1 Class Descriptions

A brief description of the categories is as follows:

5.3.1.1 Inherently Negative Bio-events

This class constitutes negated bio-events in which the event-trigger is itself a negation cue, like *independent*, *immobilization*, *unaffected*, *dysregulation*, etc. As an example, consider the sentence shown in Figure 5. The event E1 is triggered by the word *infection* and represents the initiation of viral infection of *HIV-1*. The event E2 is triggered by the word *dysregulation* and expresses the non-existence of the regulation of *Cytokine* caused by E1; therefore it has been annotated as a negated event. Similarly, the sentence in Figure 6 contains two inherently negated bio-events centred on the word *independent*. These events (E1 and E2) indicate that *the pathway in epithelial cells* is not regulated by *ROI-LOX* and *5-LOX* respectively. However, since an explicit trigger for the *Regulation* event is not present, the word *independent* has been annotated as the event-trigger.

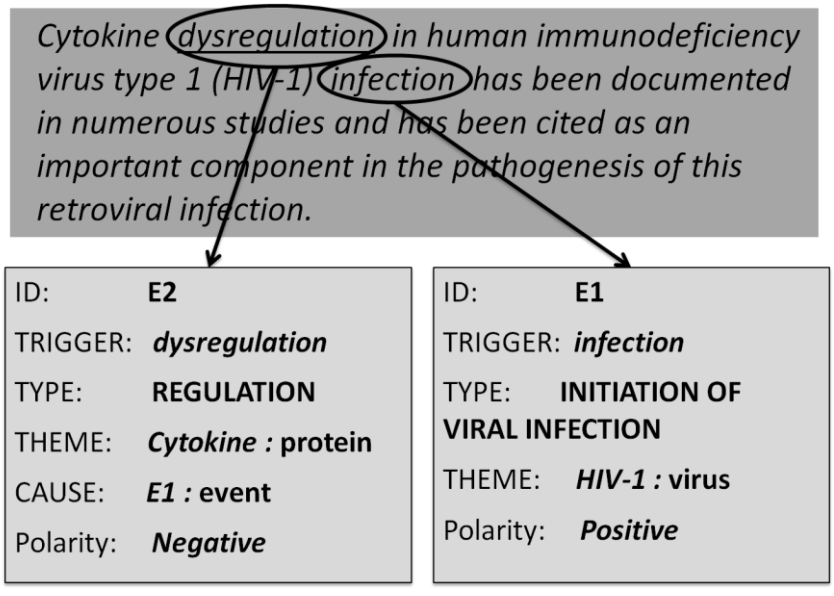

Figure 5. Inherently negative bio-event – Example 1; Source = PMID: 9427533

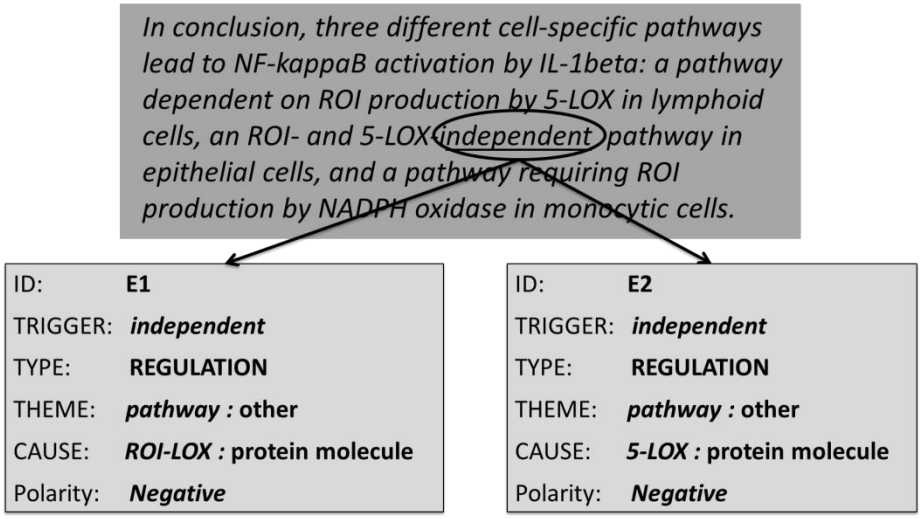

Figure 6. Inherently Negative Bio-event – Example 2; Source = PMID: 10022882

5.3.1.2 Negated Event-trigger

This class comprises bio-events in which an explicit negation cue modifies the event-trigger. For example, consider the sentence shown in Figure 7. The event E1 indicates the *Positive Regulation* of *NF-KappaB* by *IL-1beta*, where the events E2 and E3 indicate the *Regulation* of E1 by the *GTPases* (protein molecules) *Rac1* and *Cdc42*, respectively. Both E2 and E3 are negated, as they are both triggered by the word *required*, which is being modified by the explicit negation cue *not*. Interestingly, the scope of the negation cue (*not*), according to the BioScope annotation guidelines, also includes the trigger for event E1 (which is not negated). Similarly, the explicit negation cue *lacked* modifies the event-trigger for E1 in the sentence shown in Figure 8.

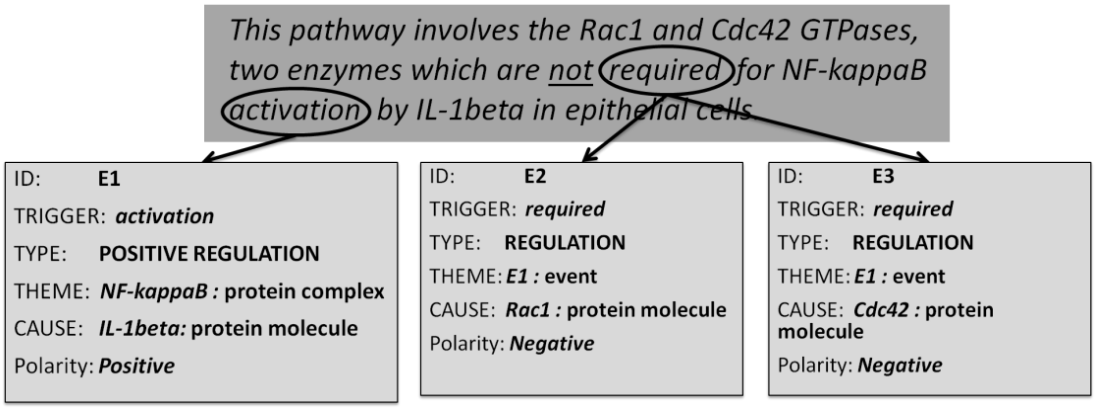

Figure 7. Negated event-trigger – Example 1; Source = PMID: 10022882

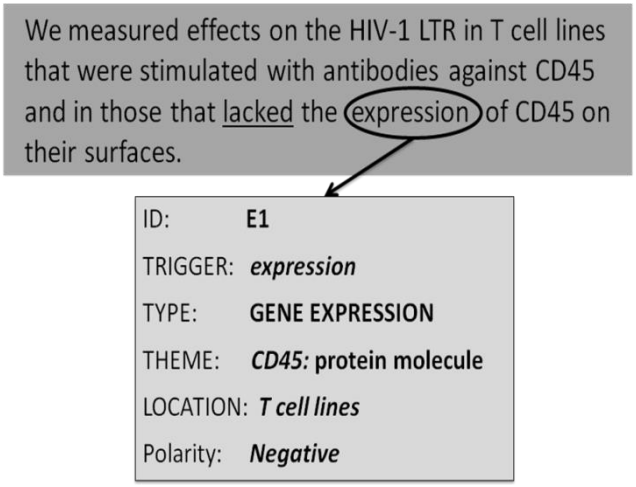

Figure 8. Negated event-trigger – Example 2; Source = PMID: 790554

5.3.1.3 Negated Participant

This class accounts for those bio-events which have at least one participant (theme or cause) being modified by an explicit negation cue. As an example, consider the sentence shown in Figure 9. Both events, E1 and E2, are triggered by the phrase *synergistically induced*; however, they have opposite polarities. Event E1 expresses the *Positive Regulation* of *IRF-1* by *IL-2* and *IL-12*, while E2 expresses the nonexistence of *Positive Regulation* of *IRF-1* by *IFN-alpha* and *IL-12*. The explicit negation cue *not* modifies the two causes of E2, i.e., *IFN-alpha* and *IL-12*.

It is important to point out that with deeper syntactic analysis, this event (E2) can instead be identified as belonging to the Negated Event-trigger class. However, since our categorisation is based on simple relationships between negation cues and individual event constituents, we have categorised such examples as instances of the Negated Participant class.

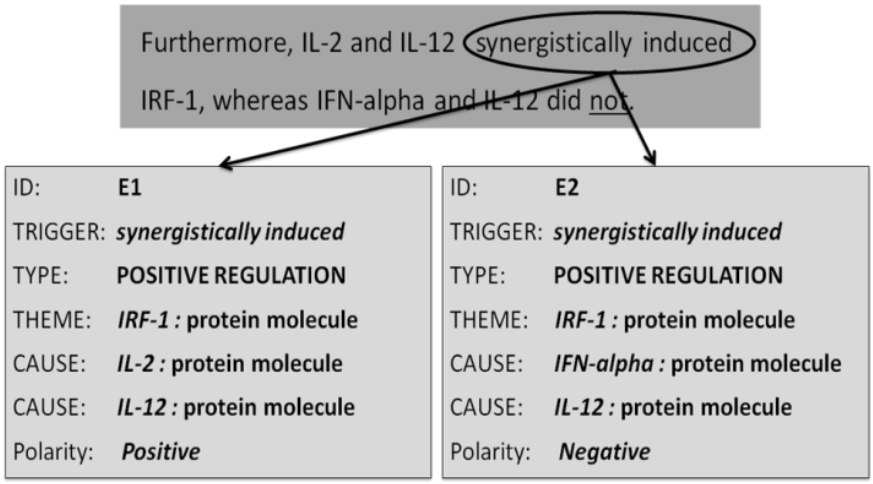

Figure 9. Negated participant; Source = PMID: 10358173

5.3.1.4 Negated Attribute

This class covers those cases of negated bio-events where an explicit negation cue modifies an event attribute, such as the location of the event. An example of this type of negation is shown in Figure 10. The events E1, E2, E3, E4, E5 and E6 are all triggered by the word *coexpressed*. However, E1 and E4 represent the *expression* of the genes *5-LOX* and *FLAP* (respectively) in *lymphoid cells*, while E2, E3, E5 and E6 represent the *expression* of these genes in *monocytic* and *epithelial cells* respectively. The explicit negation cue *not* modifies the phrase *in monocytic or epithelial cells*. This phrase contains the location for E2, E3, E5 and E6, making these events negated.

Despite its relatively low frequency, this is an important class of negated bio-events. In a recent article on the biologists' perspective of negation, Krallinger [131] identified events with negated locations as being of particular interest to biomedical practitioners.

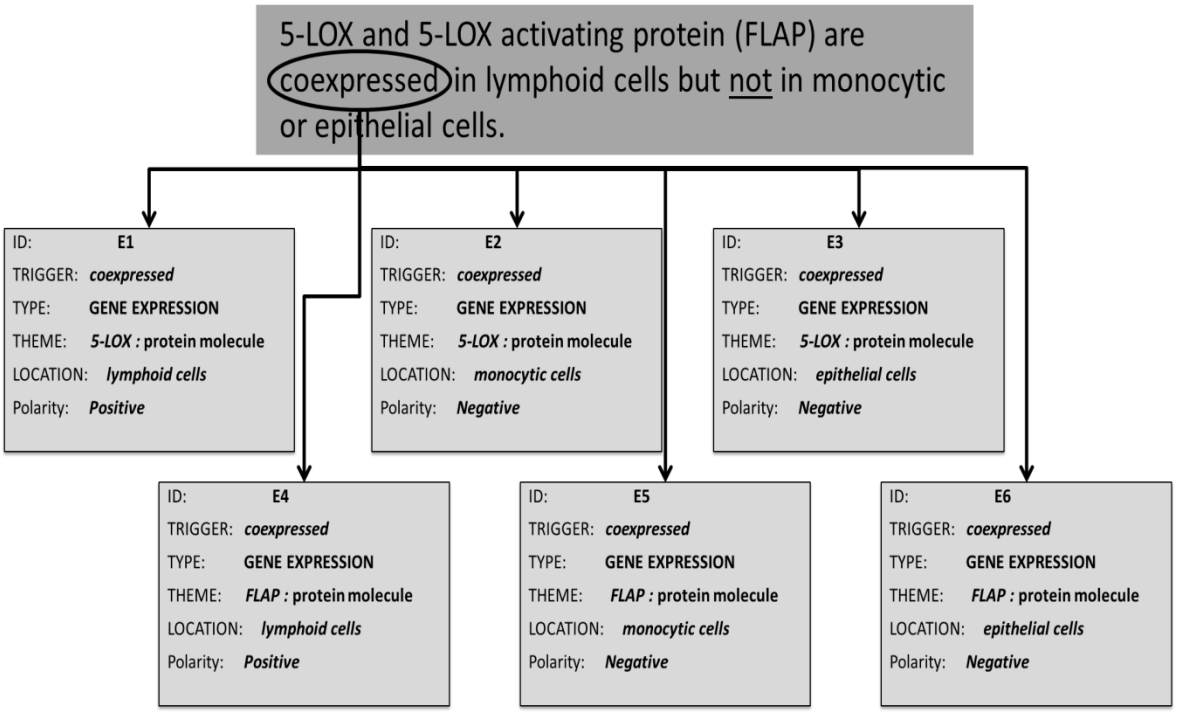

Figure 10. Negated attribute; Source = PMID: 10022882

5.3.1.5 Comparison and Contrast

This class encompasses bio-events in which a negated bio-event is signalled via contrast or comparison, normally with another bio-event. Such negated events lack an explicit negation cue. However, the BioInfer corpus is unique in the sense that it annotates even contrast and comparison markers as negation cues. Figure 11 depicts an example sentence containing a comparison-triggered negation. Event E1 is anchored to the phrase *reduced amounts*, and it expresses the *Negative Regulation* of the protein *TFIIH* in *XP-B cells*. Event E2 is triggered by the phrase *rate was normal*, and it represents the nonexistence of the *Negative Regulation* ascribed to event E1. There

is no explicit negation cue in the sentence; instead, it is the fact that the growth rate was found to be "normal" that has been used to infer that a negated event is present.

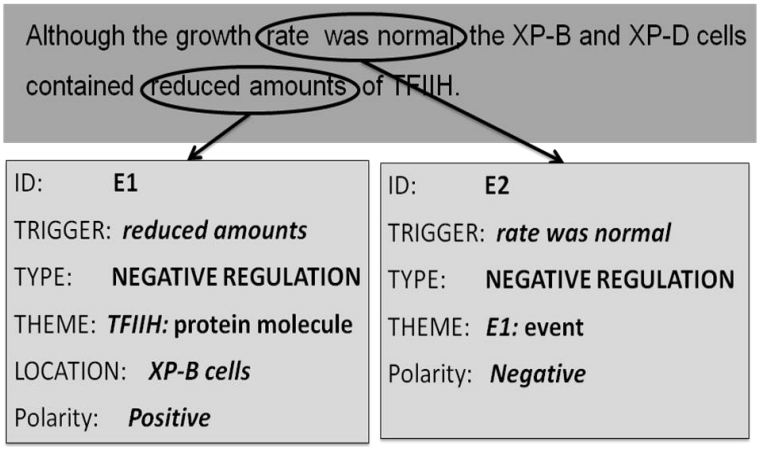

Figure 11. Comparison and contrast – Example 1; Source = PMID: 9427533

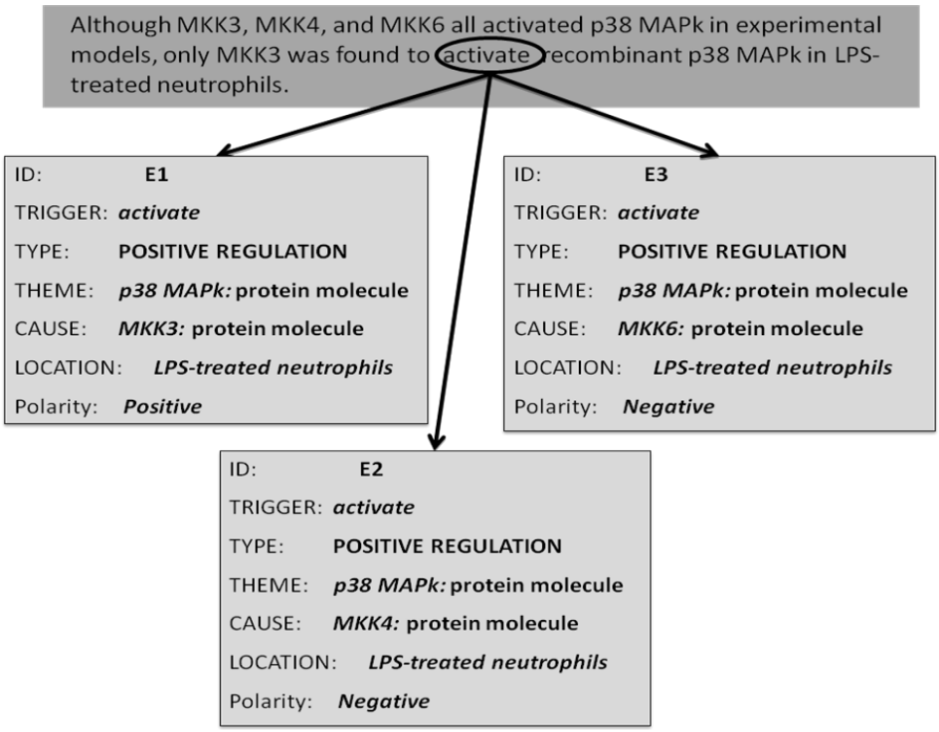

Figure 12. Comparison and contrast – Example 2; Source = PMID: 10079106

Figure 12 shows a more complex example. Event E1 is triggered by the word *activate*, and it expresses the *Positive Regulation* of *p38 MAPk* by *MKK3* in *LPS-treated*

neutrophils. Events E2 and E3 are similar to E1, except that they are not caused by *MKK3*; instead they are caused by *MKK4* and *MKK6*, respectively. Both E2 and E3 are negated; this is despite the fact that the sentence lacks an explicit negation cue.

5.3.2 Class Distribution

Our analysis revealed that the instances of each of the five classes of negated bio-events are present in the three corpora with varying frequencies. Table 21 shows the class distributions for the three corpora and the macro and micro averages for each class.

The frequency of *inherently negative* bio-events ranges between 9% and 13%, with a micro average of 12%. This is the second most prevalent category in GENIA Event and the third most prevalent category in BioInfer and BioNL'09 ST. The frequency of *negated trigger* events ranges between 61% and 67% in the three corpora, with a micro average of 63%. This is the predominant category in all three corpora. The frequency of the *negated participants* category ranges between 10% and 17%, with a micro average of 11%. This is the second most prevalent category in BioNLP'09 ST and BioInfer and the third most prevalent category in GENIA. On average, 6% of negated events belong to the *negated attribute* category; however, the frequency within the different corpora ranges between 2% and 7%. We noted that the BioInfer corpus does not mark temporal or spatial attributes of bio-events. Instead, it incorporates specialised event-types for capturing this type of information. However, some other bio-event corpora, which lack polarity information e.g. GREC, do have explicit location information. Finally, the *comparisons and contrasts* category accounts for 8% of negated bio-events.

5.3.2.1 Discussion

Previous work on the identification of negated events has primarily been focused on the *negated trigger* class, i.e., the cases where a negation cue modifies the event-trigger. However, our analysis shows that a significant proportion (37%) of negated events belongs to the other classes. Therefore, a system for effectively identifying negated bio-events should have the ability to recognise all classes of negated events.

Class	GENIA Event	BioInfer	BioNLP'09 ST	Macro Average	Micro Average
Inherently Negative	13%	11%	9%	11%	12%
Negated Trigger	61%	62%	67%	63%	63%
Negated Participant	10%	17%	12%	14%	11%
Negated Attribute	7%	2%	6%	4%	6%
Comparison and Contrast	9%	8%	6%	8%	8%

Table 21. Corpus-wise class distribution of negated bio-events

The most direct method of incorporating a particular class into a system for detecting negated bio-events is to engineer features corresponding to that class, e.g., features based on constituency or dependency relations between the negation cue and the event constituents (triggers, participants and attributes). However, features involving negation cues can only be useful for the first four classes. Since the manifestations of the *comparison and contrast* class usually lack an explicit negation cue, a different approach will be required for this class. One possibility would be to identi-

fy the comparison and contrast patterns and engineer features based on these patterns.

5.4 Key Aspects of Negation Detection – Dimensions of Analysis

Based on our analysis, we identified the key aspects of the problem of detecting negated bio-events. The intention was to use the identified aspects as dimensions of analysis for subsequent experiments, and measure the influence of each aspect on system performance.

Two obvious factors influencing any machine learning approach are the choice of features and learning algorithms. Since a vast majority of negated bio-events (over 92%) is triggered by an explicit negation cue, the choice of an appropriate list of negation cues is also very important. These negation cues can be used for subsequent feature engineering. The rest of this section discusses these three key aspects in more detail.

5.4.1 Negation Cues

Although the context and syntactic structure of the sentence play an important role in determining the negation status of a bio-event, the presence of a negation cue in the sentence is the most important factor to be considered. We define a negation cue as ‘a text fragment which causes an event to be negated’. Negation cues are usually words or phrases and they can either precede or follow the words they influence/modify [126].

5.4.1.1 Ambiguity of Negation Cues

Negation cues can be ambiguous [155, 156], i.e., in some contexts they may not trigger negations. Wilson, Wiebe and Hoffmann [157] pointed out the difference between the lexical and contextual polarities of a word. The **lexical polarity** is the prior or fixed polarity ascribed to a word, based on its meaning and general use in the language. The **contextual polarity** of a word is more dynamic and depends on the context of the text fragment containing the word. The contextual polarity can be different from the lexical polarity, and this difference is the key source of ambiguity in determining the negation cue status of a word or phrase. For example, consider the words *lack* and *loss*. Both of these words have a negative lexical polarity, as they convey the "state of not having something". That is why they have been identified as negation cues in the BioScope corpus. Morante [155] also identified both of these words as unambiguous negation cues. However, from a biological perspective, these words have a positive polarity when used in the context of a *negative_regulation* event. Hence, a positive contextual polarity can be ascribed to these words in certain instances. Similarly, the words *absent* and *absence* may also be used to convey *negative regulation*, rather than negation.

Previously, in Figure 8, we showed a case of the word *lacked* acting as a negation cue. This is a case of matching lexical and contextual polarities. However, Figure 13 shows a case of conflicting lexical and contextual polarities. In the sentence shown, the event E1 is anchored to the word *loss*, and it expresses the *negative_regulation* of the protein molecule *STAT1* in *cells from patients treated with fludrabine in vivo*. In this case the polarity of E1 is positive.

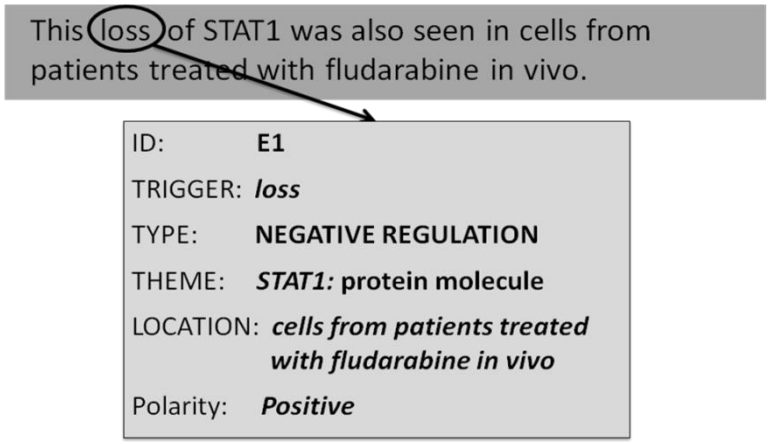

Figure 12. An instance of the word *loss* with positive contextual (biological) polarity; Source = PMID: 10202937

Our analysis of negated bio-events has led us to the conclusion that the ambiguity status of a negation cue is not universal. Instead, it is determined by the:

- Nature of text under consideration
- Annotation perspective (e.g., linguistic or biological)
- Textual context and lexical polarity of the cue

5.4.1.2 Indicators of Low Manner of Interaction

Sometimes, the text containing a bio-event also contains a word or phrase which provides an indication of the rate, level, strength or intensity of the interaction. As explained in chapter 3, we refer to this indication as the *manner* of the event [8], and distinguish between three types of manner: *high*, *neutral* and *low*. The words indicating a low manner include adjectives and adverbs like *weak*, *weakly*, *slight*, *slightly*, *slow*, *small*, *little*, *low*, etc.

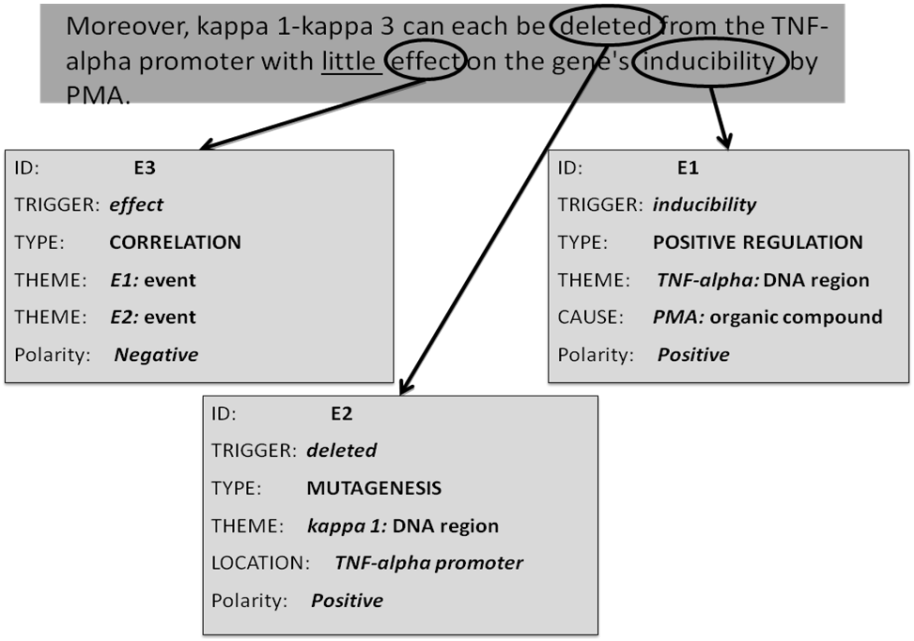

Figure 13. An instance of the low manner indicator little being treated as a negation cue; Source = PMID: 20562282

Indicators of low manner have historically been treated as negation cues. In the field of sentiment analysis, the indicators of low manner have been considered a special class of negative polarity indicators. Wiegand et al. [156] refer to this class of cues as the *diminishers*, while Wilson, Wiebe and Hoffmann [157] labelled them as *negative polarity shifters*. Similarly, indicators of low manner have been treated as negation cues in the field of biomedical text mining. Examples include the three corpora of negated bio-events (i.e., GENIA Event, BioInfer and BioNLP'09 ST) and the BioScope corpus. Figure 14 shows an example sentence where the low manner indicator *little* has been interpreted as a negation cue for the event E3.

In our model of event interpretation [8], polarity and manner are treated as orthogonal dimensions of event interpretation, i.e., the value of manner does not influence

the value of polarity and vice-versa. According to this approach, the event E3 in Figure 14 will have a *low* manner but a *positive* polarity.

5.4.1.3 Deactivators of Negation Cues

The capacity of some words to act as negation cues is affected by the constructions in which they are used. This means that a word that normally acts as a negation cue can cease to act in that way if it is preceded and/or followed by certain other words. We refer to these syntactic patterns as *negation deactivation patterns*. Here, we focus only on the two most common negation cues, i.e., *no* and *not*.

Deactivators of Not

The word *not* is the most frequent negation cue in the BioScope corpus, where it accounts for over 41% of the total negation instances. However, in almost 8% of cases, it does not indicate a negation, i.e., it ceases to act as a negation cue. In our analysis, we focused on a simple deactivation pattern: *not* <*deactivatorOfNot*>. The pattern indicates an occurrence of the word *not* immediately followed by one of its *deactivators*. We only considered the following five deactivators: *clear*, *evident*, *known*, *necessarily* and *only*.

In our analysis of the GENIA Event corpus, we discovered a total of 261 events which belonged to the sentences containing the above pattern. Amongst these, 258 events (99%) were positive and only 3 events (1%) were negated, suggesting that this is an effective pattern to identify the deactivated instances of the word *not*.

Deactivators of No

The word *no* is the second most frequent negation cue in the BioScope corpus and accounts for almost 30% of the total negation instances in the corpus. However, in

over 6% of cases, it does not indicate a negation. Morante [155] has identified some constructions which contain the word *no*, but do not trigger a negation. These constructions include: *no sign of*, *no evidence of*, *no proof* and *no guarantee that*, etc.

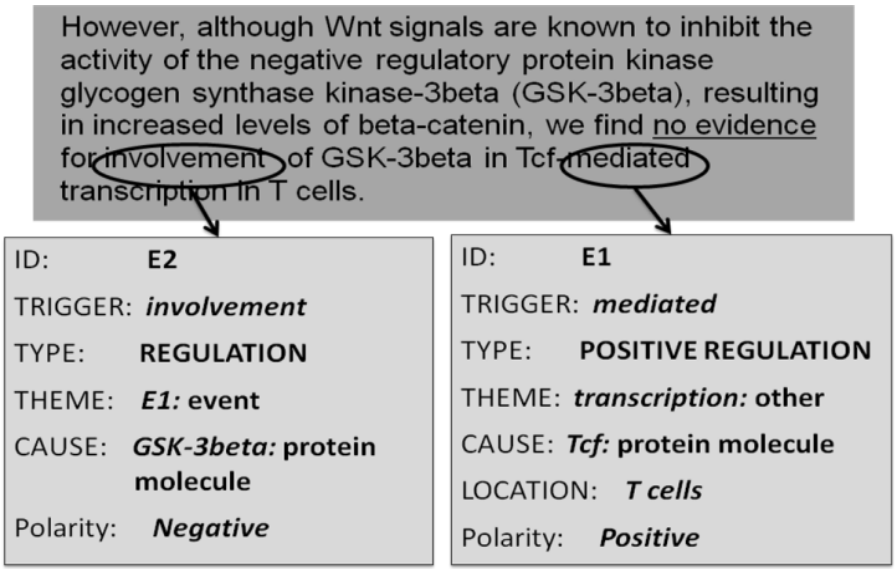

Figure 14. An instance of negation triggered by the construction *no evidence*; Source = PMID: 10221643

Our analysis of the GENIA Event corpus revealed that in some cases, these constructions do trigger negated events. For example, consider the sentence in Figure 15, where the construction *no evidence* triggers the negation of event E2. Based on our analysis, we conclude that the deactivation patterns identified for linguistic (scope) annotation may not hold for biological (event) annotation.

5.4.1.4 Relationship between Negation Cues and Event-types

We investigated the relationship between negation cues and different types of bio-events. Our analysis revealed two classes of negation cues with respect to event-types. These are:

Type-independent Negation Cues

This class includes the typical negation markers like *no*, *not* and *fail*, etc. Some inherently negative event-triggers which can be applied to various types of events are also included in this category. For example, event-triggers like *unaffected* and *independent* can be used for various types of events including *Positive Regulation*, *Negative Regulation* and *Correlation* events.

Type-dependent Negation Cues

This class includes cues like *immobilize*, *decoupling* and *dysregulation*, which act as negation cues for specific event-types only: *immobilize* and *decoupling* for *Localization* events only and *dysregulation* for *Regulation* events only.

5.4.1.5 Corpus / Domain Idiosyncrasies

Some cues which are unambiguous and/or frequent in one corpus can be ambiguous and/or scarce in another. For example, words like *protected* and *abolish* are treated as negation cues in BioInfer. However, they are mostly interpreted as indicators of *Negative Regulation*, rather than negation, in the GENIA Event and BioNLP'09 ST corpora.

In contrast, the verb *fail* is frequent and mostly unambiguous in the GENIA Event and BioNLP'09 ST corpora. However, in the BioInfer corpus, it does not appear as a negation cue even once.

5.4.1.6 Compilation of Cue Lists

Having identified negation cues as an important factor for the identification of negated bio-events, we conclude that it is important to:

- determine the impact of the choice of cue lists on the overall task performance
- identify an optimum cue list for the task

Based on the above analysis, we decided:

1) not to create separate lists for ambiguous and unambiguous cues
2) to treat the *low manner* indicators as negation cues

We then compiled four separate lists of negation cues for comparison. Table 22 depicts the elements in each list. A brief description of each list is as follows:

c40

We formulated a list of 40 cue words by combining previously published lists and cues discovered during our own initial analysis of negated bio-events. We did not include any phrases in the list.

cBioInfer

We extracted the negation cues from the BioInfer corpus. This was a straightforward task, because the cues had already been annotated. We then selected the top 25 cue words (no phrases) to form the cBioInfer list.

cBioScope

This is the list of 28 negation cues, including both words and phrases, compiled by Morante [155] from the BioScope corpus.

Name	Size	Elements
c40	40	absence, absent, barely, cannot, deficiency, deficient, except, exception, fail, failure, impair, inability, inactive, independent, independently, insensitive, instead, insufficient, lack (noun), lack (verb), limited, little, loss, lose, lost, low, negative, neither, never, no, none, nor, not, prevent, resistance, resistant, unable, unaffected, unchanged, without
cBioScope	28	absence, absent, cannot, could not, either, except, exclude, fail, failure, favor over, impossible, instead of, lack (noun), lack (verb), loss, miss, negative, neither, never, no, no longer, none, not, rather than, rule out, unable, with the exception of, without
cBioInfer	25	abolished, absence, cannot, defective, deficient, despite, differ, different, differential, distinct, failure, independent, independently, lack, negligible, neither, no, nor, not, protected, separately, simultaneously, unable, unlike, without
cCore	19	absence, fail, inability, independent, independently, insensitive, insufficient, lack (noun), lack (verb), little, neither, no, nor, not, resistant, unable, unaffected, unchanged, without

Table 22. Cue lists

cCore

As previously discussed in section 5.3, we analysed 1,000 randomly selected negated bio-events (600 from GENIA Event, 300 from BioNLP'09 ST and 100 from BioInfer). We made a list of all negation cues observed in these bio-events, and conducted a series of experiments to identify the smallest set of cues for optimum performance. Based in these experiments, we compiled the cCore cue list which contains only 19 cue words.

5.4.2 Feature Design

Feature engineering and selection is a vital part of any machine learning system. Various types of features have previously been used for different negation detection tasks, including lexical, syntactic, semantic and statistical (bag of words) features. However, most previous work on detection of negated bio-events has concentrated around event-triggers, whilst the other semantic aspects of the event (like location and participants) have been ignored.

The aim of our investigation was to:

- identify the optimum set of features for the task of identifying negated bio-events
- compare the performance of different feature sets by evaluating their individual and combined impact on the overall performance of a system for detecting negated bio-events

In order to achieve the above aims, we used our analysis of negated bio-events to engineer various semantic, lexical and syntactic features. Using this preliminary set of features, we conducted a series of experiments to identify the minimum optimal

feature set required for the task of identifying negated bio-events. Table 23 shows the 17 features that form this optimum set according to our experiments. These features have been grouped into four categories; we have labelled these sets as Semantic, Lexical, Dependency and Command.

Feature Set	ID	Value Function
Semantic	S1	isComplex(event)
	S2	eventType(event)
Lexical	L1	contains(sentence, negCue)
	L2	negCue()
	L3	isNextTo(negCue, deactivator)
	L4	minimum(distance(eventTrigger, negCue), distance(eventLocation, negCue))
	L5	contains(eventTrigger, negCue)
Dependency	D1	relation(negCue, eventTrigger) || relation(negCue, eventLocation)
	D2	relation(negCue, eventTheme) || relation(negCue, eventCause)
	D3	relation(negCue, X) && (relation(X, eventTrigger)||relation(X, eventLocation))
	D4	relation(negCue, X) && (relation(X, eventTheme)||relation(X, eventCause))
Command	C1	sCommands(negCue, eventTrigger) || sCommands(negCue, eventLocation)
	C2	sCommands(negCue, eventTheme) || sCommands(negCue, eventCause)
	C3	vpCommands(negCue, eventTrigger) || vpCommands(negCue, eventLocation)
	C4	vpCommands(negCue, eventTheme) || vpCommands(negCue, eventCause)
	C5	npCommands(negCue, eventTrigger) || npCommands(negCue, eventLocation)
	C6	npCommands(negCue, eventTheme) || npCommands(negCue, eventCause)

Table 23. Feature sets

5.4.2.1 Semantic Features

The purely semantic features are constructed from the semantic information available for the bio-event. This information includes the semantic type of the bio-event (e.g., *gene_expression*, *localization*, *positive_regulation* etc.), the semantic type of each participant (e.g., *lipid*, *DNA molecule* and *protein complex* etc.) and the role of each participant (e.g., *theme* and *cause*, etc.). Table 23 shows the two such features which were found to be useful. Feature S1 indicates whether a bio-event is complex i.e., whether it has one or more participants which are bio-events themselves. Feature S2 is the semantic type of the bio-event.

5.4.2.2 Lexical Features

The purely lexical features are constructed from the sentence containing the bio-event. Table 23 shows three features of this type: L1, L2 and L3. Feature L1 indicates whether the sentence contains any of the negation cues from a specified list. Feature L2 is the negation cue itself; if the sentence does not contain a negation cue, then this feature is assigned a default value. Feature L3 indicates whether a specified negation deactivator is situated next to the negation cue in the sentence. This is a novel feature, which has not been used previously for negation detection tasks.

The lexico-semantic features are constructed using a combination of the "textual" bio-event information and the sentence containing the bio-event. The textual bio-event information includes the text fragment indicating the occurrence of the bio-event (i.e., the event-trigger), the text fragments identifying the event participants and the text fragments indicating any event attributes like location, etc. Table 23 shows the two lexico-semantic features which are included in the optimum set. Feature L4 is the minimum of the surface distances between the event-trigger and the

negation cue and the surface distance between the event-location and the negation cue. This is a novel feature, as none of the previous studies have considered event location. Feature L5 indicates whether a negation cue forms part of the event-trigger. This feature has been engineered to account for the class of inherently negative bio-events.

5.4.2.3 Dependency Features

These are the lexico-syntactic features constructed using the textual bio-event information and the dependency relations found in the sentence. All of these features are novel, and they have been especially engineered to incorporate specific classes of negated bio-events discussed in section 5.3. Unlike certain previous studies [132, 134], we have not based the features on specific dependency relations. Instead, the existence/non-existence of any dependency relation between specific text fragments has been used as the basis for these features.

Table 23 shows four dependency features. Feature D1 indicates whether there is a dependency relation between the negation cue and the event-trigger, or between the negation and the event-location. Feature D2 indicates whether there is a dependency relation between the event-trigger and the event-theme, or between the event-trigger and the event-cause. Features D3 and D4 are more complex. D3 indicates whether there is an indirect (single hop) dependency relation between the negation cue and the event-trigger or the event-location. Similarly, D4 indicates whether an indirect relation exists between the event-trigger and event-theme or event-cause.

5.4.2.4 Command Features

These are the lexico-syntactic features constructed from the textual bio-event information and the *command relations* found in the constituency parse tree of the sentence. The concept of a command relation was first introduced by Langacker [158] as a means for identifying the nodes affected by a given element in the constituency parse tree of a sentence. He defined a command relation as follows: 'a node *X* commands a node *Y* if neither *X* nor *Y* dominates the other and the *S* (sentence) node most immediately dominating *X* also dominates *Y*'. Reinhart [159] introduced the more general concept of *constituent command* which is often abbreviated *as c-command*. She defined the c-command as follows: 'node *X* c-commands node *Y* if neither *X* nor *Y* dominates the other and the first branching node that dominates *X* also dominates *Y*'. Baker and Pullum [160] relaxed the definition of a command relation by eliminating the mutual non-dominance condition and relabelled it as the *S-command* relation. Their definition of S-command is as follows: 'a node *X* S-commands a node *Y* if the *S* node immediately dominating *X* also dominates *Y*'.

We have engineered three types of command features using the generic *Q-command* relation. We define the Q-command relation as follows: 'a node *X* Q-commands node *Y* if the first dominant Q node of *X* also dominates *Y*'. We use three types of command relations (i.e., three values of Q): S-command, VP-command and NP-command.

Table 23 includes six novel command features, covering specific classes of negated bio-events discussed in section 5.3. Feature C1 indicates whether the negation cue S-commands either the event-trigger or the event-location. C2 indicates whether the negation cue S-commands either the event-theme or the event-cause. Features C3

and C4 are similar to C1 and C2; however, they are based on the VP-command relation. Similarly, features C5 and C6 are based on the NP-command relation.

5.4.3 Choice of Learning Algorithm

The choice of learning algorithm can significantly influence the performance of a classification task. This has been demonstrated for various natural language processing tasks including text categorization [161], word sense disambiguation [162] and the detection of negated terms [151]. In order to measure the impact of the choice of learning algorithm on the task of identifying negated bio-events, we decided to compare the performance of the most commonly used learning algorithms. We selected the following six algorithms for this task:

5.4.3.1 Decision Trees

Decision Tree algorithms learn rules which are expressed as "conjunctions of constraints on the attribute values of instances. Each path from the tree root to a leaf corresponds to a conjunction of attribute tests, and the tree itself to a disjunction of these conjunctions" [163]. Various Decision Tree algorithms have been proposed over the years. However, we concentrated on C4.5 [164], which is an enhanced version of ID3 [165]. The C4.5 algorithm constructs the Decision Tree by choosing the attribute with the highest value of normalised information gain at each node, and creates new branches corresponding to the different values of this attribute. Once the initial tree has been created, the algorithm tries to identify and remove the least useful branches. Decision trees have been extensively used for various problems in bioinformatics [166]. They have also been used to detect negations in medical texts [150].

5.4.3.2 Random Forest

The Random Forest [167] algorithm develops an ensemble (i.e., a forest) of Decision Trees from randomly sampled subspaces of the input features. Once the forest has been created, new objects are classified using a two-step process:

1) An individual classification is obtained from each tree in the forest.

2) The final classification of the object is determined by majority votes among the classes obtained from individual trees.

Despite being successfully used for various text mining and bioinformatics tasks [168, 169], the Random Forest algorithm has not been previously used for detecting negation scopes, negated concepts or negated events.

5.4.3.3 Logistic Regression

Logistic Regression classifiers try to predict the class probability of an object by fitting the training data to a logistic function. Logistic Regression classifiers have previously been used to identify negated bio-events [134].

5.4.3.4 Naive Bayes

Naïve Bayes is one of the simplest probabilistic classification algorithms. It uses the Bayes probability model for predicting the class probabilities of inputs. The word *naïve* indicates that the algorithm assumes class conditional independence i.e., it assumes that the effect of a variable value on a given class is independent of the values of other variable. Despite its simplicity, the Naïve Bayes algorithm achieves good results for many complex classification problems [170]. It has also been used to detect negations in medical texts [150, 151].

5.4.3.5 SVM

Support Vector Machines (SVM) [171] perform classification by constructing an *N*-dimensional hyperplane that optimally separates the data into two categories. They use a kernel function to transform the data into a higher dimensional space, which paves the way for optimal separation. Many previous studies in negation detection have used SVM [136, 139, 151].

5.4.3.6 Instance-Based Algorithms

The Instance-Based (also known as Memory-Based) learning algorithms do not derive generalisations or abstractions from the complete training data. Instead, they keep all training data in memory, and generate classification predictions using only the most similar training instances. IB1 [172] is an instance-based learning algorithm. It uses normalised Euclidean distance to find the training instance closest to the given test instance, and predicts the same class as this training instance. IB1 is similar to the nearest neighbour algorithm, except that it normalises its attributes' ranges, processes instances incrementally, and has a simple policy for tolerating missing values. Instance based learning algorithms have previously been used for detecting negation cues and their scopes [139].

5.5 Experimental Settings

This section presents a brief description of the experimental set-up, including the datasets, parsers, classifiers and the evaluation metrics.

5.5.1 Datasets

We performed experiments using all three open access corpora of negated bio-events. These corpora were discussed in section 5.1.1.

5.5.2 Parsing

We used the Enju parser [173] to extract the POS tags, phrase structure trees and dependency relations. Enju is a deep parser which uses a Head-driven Phrase Structure Grammar (HPSG) extracted from the Penn Treebank and a maximum entropy model trained with an HPSG tree-bank derived from the Penn Treebank. It achieves a parsing accuracy of around 90% on both newswire articles and biomedical papers. Enju presents the parsing output in the form of a predicate-argument structure, which is a graph structure that represents syntactic/semantic relations among words. The Enju output also includes predicate-argument relations, which are the dependency relations between pairs of words.

5.5.3 Classifier Implementation

We used the WEKA [174] library for constructing our classifiers. The implementation details for each algorithm are as follows:

- The C4.5 implementation in WEKA is based on [164]. We used the following optimisation settings: (1) apply sub-tree replacement, (2) apply sub-tree raising, (3) require a minimum of 2 instances per leaf, (4) set a confidence threshold for pruning of 0.25.
- The Random Forest implementation in WEKA is based on [167]. Our optimisation settings included: (1) set the number of trees in the forest to 10, (2) set the number of features used to build individual trees to log(N+1), where N is the total number of features, (3) set no restrictions on the depth of individual trees.

- The Logistic Regression implementation in WEKA is a slightly modified version of [175]. No optimisation settings were used.
- The WEKA implementation of the Naïve Bayes algorithm uses a default precision of 0.1 for numeric attributes for cases of zero training instances. We used the default settings.
- The SVM implementation in WEKA is based on the sequential minimal optimisation algorithm by Platt [176]. This implementation replaces all missing values, and converts the nominal attributes to binary attributes. It also normalises all attributes by default. We used: (1) a polynomial kernel, (2) the default value of the complexity constant.
- The WEKA implementation of the IB1 algorithm is based on [172]. We used the default settings.

5.5.4 Evaluation Measures

We used the standard metrics of precision, recall and F-measure for reporting and comparing results. Precision is the number of true positives divided by the sum of true positives and false positives; recall is the number of true positives divided by the sum of true positives and false negatives; and F-measure is the first harmonic mean of precision and recall. These metrics are regularly used to report results for various text mining tasks [2].

5.6 Results

We ran a series of experiments for each dataset to systematically evaluate the impact of each of the four cue lists, the six learning algorithms and the four main feature

sets and their combinations. This section describes the results of our experiments. All results are based on 10-fold cross validation.

5.6.1 Best Results for Each Dataset

On the GENIA Event dataset, the best results were achieved using the Random Forest classifier using all four feature sets engineered from the c40 cue list. The classifier achieved 83% precision and 67% recall, leading to an F-score of 74%. The same classifier achieved the best results on the BioNLP'09 ST dataset, achieving approximately 78% precision, 64% recall and 70% F-score. The best results on the BioInfer dataset were also achieved by a Random Forest classifier with all feature sets; however, the cBioInfer cue list was used to engineer the features. This classifier achieved 86% precision, 85% recall and 85% F-score. Table 24 shows the best results achieved for each dataset.

Dataset	P	R	F	Algorithm	Cue List	Features
GENIA Event	83.1%	67.1%	74.2%	Random Forest	c40	All
BioInfer	86.1%	84.5%	85.3%	Random Forest	cBioInfer	All
BioNLP'09 ST	77.6%	63.9%	70.1%	Random Forest	c40	All

Table 24. Best results for each dataset

5.6.2 Cue List Comparison

In order to compare the performance of the four cue lists, we ran a series of experiments using the Random Forest algorithm. We chose the Random Forest algorithm because it had consistently produced the best results for all datasets. For each dataset, we constructed a Random Forest classifier using all four feature sets. Howev-

er, the cue list used to engineer the features was varied. Table 25 shows the performance of the four cue lists for each of the datasets. The key results are as follows:

- The c40 cue list performed well on all three datasets. It outperformed the other cue lists on GENIA Event and BioNLP'09 ST, and achieved the highest precision, recall and F-score on both datasets. However, on BioInfer it performed worse than cBioInfer and cCore.
- The cCore cue list performed consistently, and achieved the second best results (F-score) for all three datasets. Its results were very close to the top performing cue list for GENIA Event and BioNLP'09 ST with margins of 0.5% and 1.8%, respectively. However, on BioInfer it was second by a significant margin of 7%.
- The cBioInfer cue list lagged behind c40 and cCore by almost 5% and 8% on GENIA Event and BioNLP'09 ST. However, as expected, it achieved the best results on BioInfer by a fair margin (over 7%).
- The cBioScope cue list achieved the lowest results for all three datasets by significant margins (ranging between 6% and 8%).

Cue List	GENIA Event			BioInfer			BioNLP'09 ST		
	P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
c40	**83.1%**	**67.1%**	**74.2%**	84.4%	70.8%	77.0%	**77.6%**	**63.9%**	**70.1%**
cCore	82.6%	66.7%	73.8%	87.0%	70.8%	78.1%	76.3%	61.6%	68.2%
cBioInfer	81.4%	60.4%	69.3%	86.1%	**84.5%**	**85.3%**	75.3%	53.2%	62.3%
cBioScope	80.7%	59.9%	68.8%	**89.3%**	67.7%	77.0%	75.4%	52.9%	62.2%

Table 25. Cue list comparison

Figure 16 shows the micro-averaged results for each cue list. It shows that overall (in terms of F-score) c40 performed the best, followed by cCore (-0.7%), cBioInfer (-4.8%) and cBioScope (-5.7%), respectively.

The difference between the best and the worse performance caused by the choice of cue list was 5% for GENIA Event, 7% for BioInfer and 8% for BioNLP'09 ST. This provides sufficient evidence in favour of the hypothesis that the choice of the negation cues used for engineering the feature set has a significant impact on performance of a system designed for the identification of the negated bio-events.

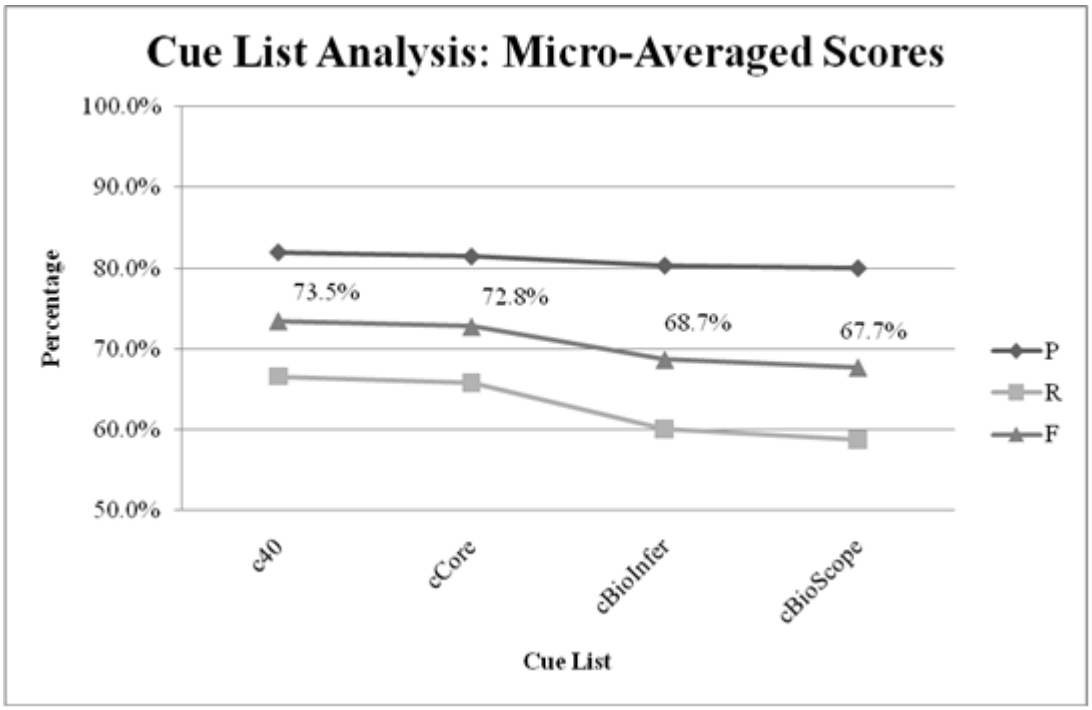

Figure 15. Cue list comparison: Micro-averaged results for the three datasets

5.6.3 Feature Set Comparison

Gain Ratio

We computed the gain ratio, of both individual features and feature sets, on the three datasets. The dependency features achieved the highest gain ratio, followed by the

command and lexical features, respectively. The semantic features achieved the lowest gain ratios. In terms of individual features, D1 achieved the highest score, by a fair margin, for all three datasets. Features D4, C3 and L4 also achieved consistently high scores on all three datasets.

Classification Results

In order to compare the performance of the various features, we ran a series of experiments on each dataset. In each experiment, we constructed a Random Forest classifier using a different combination of features, which were engineered from the cCore cue list. We chose the Random Forest algorithm and the cCore cue list because both had performed consistently on all three datasets. Table 26 shows the results for the four feature sets and some of their combinations. The key findings are as follows:

- Using only the semantic features, the Random Forest algorithm could not find a model. This was mainly due to the small number of semantic features used and the relatively poor discriminative ability of these features, as evidenced by the low information gain scores.
- The lexical features achieved the highest scores as an individual feature set on GENIA Event and BioNLP'09 ST, and the second highest score as an individual feature set on BioInfer.
- The dependency features achieved the highest score as an individual feature set on BioInfer and the second highest scores on GENIA Event and BioNLP'09 ST.

- The command features scored significantly lower than the lexical and dependency features.
- The combination of lexical and dependency features outperformed the combination of lexical and command features on all three corpora.
- The combination of all four feature sets achieved the best overall results for all three datasets.

Cue List	GENIA Event			BioInfer			BioNLP'09 ST		
	P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
Sem	No model found			No model found			No model found		
Lex	83.3%	54.9%	66.2%	69.4%	53.4%	60.4%	**82.6%**	48.2%	60.9%
Dep	67.4%	59.0%	62.9%	74.1%	51.6%	60.8%	73.8%	46.8%	57.3%
Com	53.9%	14.2%	22.5%	63.7%	36.0%	46.0%	68.5%	32.0%	43.6%
Lex + Com	79.6%	57.7%	66.9%	78.8%	64.6%	71.0%	77.2%	54.3%	63.8%
Lex + Dep	**84.4%**	61.8%	71.4%	83.1%	64.0%	72.3%	79.7%	57.1%	66.5%
All	82.6%	**66.7%**	**73.8%**	**87.0%**	**70.8%**	**78.1%**	76.3%	**61.6%**	**68.2%**

Table 26. Feature set comparison

Figure 17 shows the micro-averaged results for the four feature sets and their three combinations. Overall, in terms of individual feature sets, the lexical features perform slightly better than the dependency features, while the command and semantic features perform significantly worse. The combination of lexical and dependency features performs better than the combination of lexical and command features. However, the combination of all four feature sets achieves the highest scores.

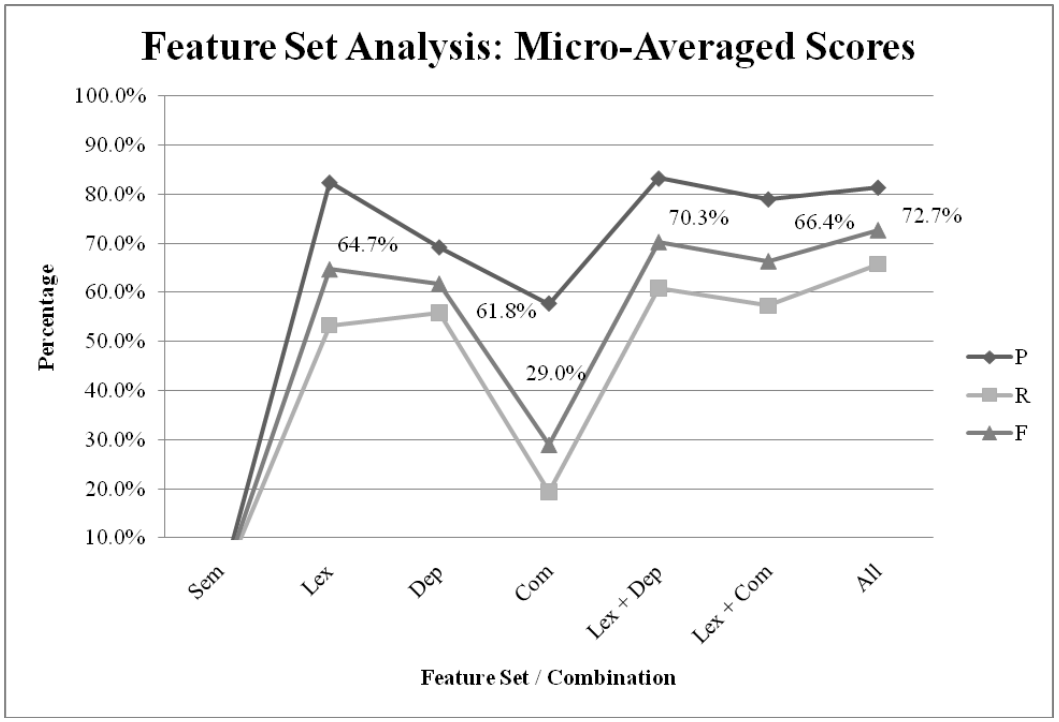

Figure 16. Feature set comparison: Micro-averaged results for the three datasets

5.6.4 Algorithm Comparison

In order to compare the performance of the chosen learning algorithms for the task of identifying negated bio-events, we ran a series of experiments on each dataset. In each experiment, we constructed a classifier using the chosen algorithm and all feature sets. The features were engineered from the cCore cue list. We chose the cCore cue list because it had performed consistently on all three datasets. Table 27 shows the results for each dataset. The key findings are as follows:

- C4.5 performed consistently on all three datasets. It outperformed the other algorithms on BioNLP'09 ST, scored second on GENIA Event and fourth on BioInfer.
- Random Forest outperformed the other algorithms on GENIA Event and BioInfer, and scored second on the BioNLP'09 ST by a narrow margin of 0.8%.

- Logistic Regression achieved the third best results on both GENIA Event and BioInfer. It scored fourth on BioNLP'09 ST.
- Naive Bayes achieved the highest recall for all datasets. However, its precision was noticeably low (ranging between 32% and 42%), which led to the lowest F-scores for all datasets.
- SVM scored fifth for all three datasets. Although it performed much better than Naive Bayes, it was significantly behind Random Forest and C4.5
- IB1 gave the second best results for BioInfer and the fourth best results for both GENIA Event and BioNLP'09 ST.

Algorithm	GENIA Event			BioInfer			BioNLP'09 ST		
	P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
C4.5	**84.4%**	62.4%	71.8%	82.1%	68.3%	74.6%	**82.2%**	56.5%	**67.0%**
Random Forest	82.6%	66.7%	**73.8%**	**87.0%**	70.8%	**78.1%**	76.3%	58.4%	66.2%
Logistic Regression	82.8%	58.7%	68.7%	79.3%	71.4%	75.1%	80.5%	53.1%	64.0%
Naïve Bayes	31.6%	**83.0%**	45.8%	42.2%	**83.9%**	56.2%	32.9%	**82.3%**	47.0%
SVM	79.3%	53.7%	64.0%	79.0%	67.7%	72.9%	78.6%	46.7%	58.6%
IB1	66.1%	66.7%	66.4%	85.8%	71.4%	77.9%	70.8%	59.5%	64.7%

Table 27. Algorithm comparison

Figure 18 shows the micro-averaged results for each algorithm. It shows that overall (in terms of F-score), Random Forest performed the best, followed by C4.5 (-1.5%), Logistic Regression (-4.3%), IB1 (-5.6%), SVM (-8.9%) and Naive Bayes (-25.7%).

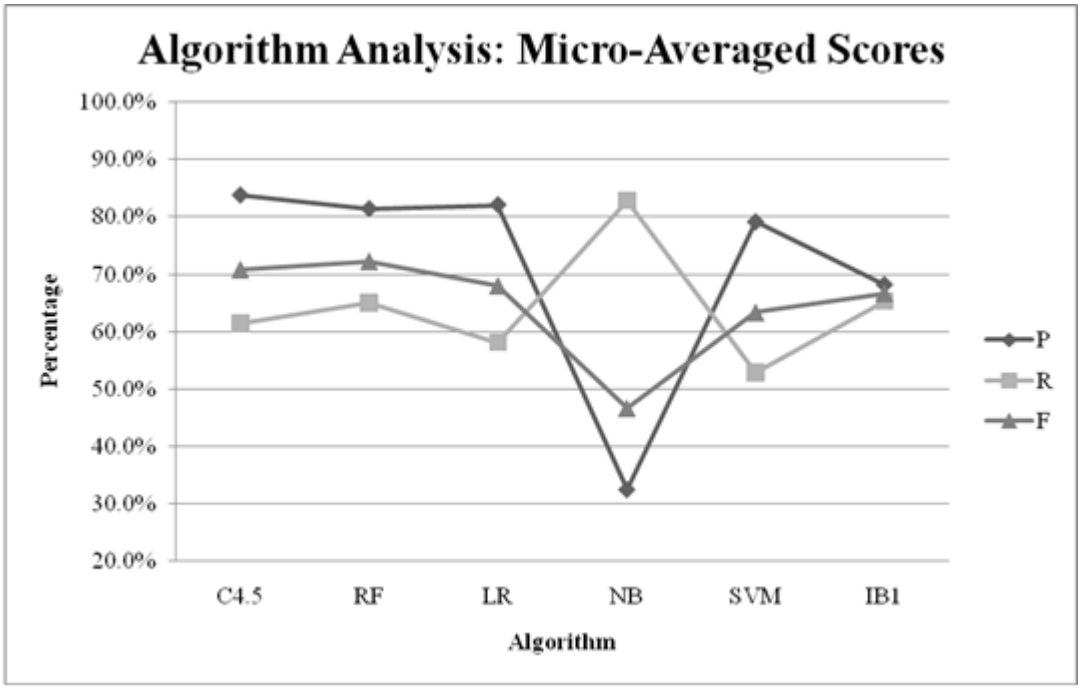

Figure 17. Algorithm comparison: Micro-averaged results for the three datasets

The difference between the best and the worst performing algorithms was 28% for GENIA Event, 22% for BioInfer and 20% for BioNLP'09 ST. Even if we exclude Naive Bayes, which performed significantly worse than the rest of the algorithms, the difference was still 10% for GENIA Event, 5% for BioInfer and 8% for BioNLP'09 ST. This provides sufficient evidence in favour of the hypothesis that the choice of learning algorithm has a significant impact on the performance of a (machine learning) system for identifying negated bio-events.

5.7 Discussion

This section provides a brief discussion on the key aspects of our results and findings.

5.7.1 Comparison with Previous Results

As mentioned earlier, the identification of negated bio-events is a new area of research and only a few results have been reported previously. The previously best reported results for identification of negated bio-events were by Sarafraz and Nenadic [136]. They used the *Training* subset of the BioNLP'09 ST dataset for training and the *Development* subset for testing. They achieved 38% precision, 76% recall and 51% F-score. In comparison, our system achieved an F-score of above 70% with 10-fold cross validation on the entire BioNLP'09 ST dataset. In order to obtain a more direct comparison, we conducted further experiments with the same experimental settings as those used by Sarafraz and Nenadic [136]. That is, we trained our Random Forest classifier on the *Training* subset of the BioNLP'09 ST data and tested it on the *Development* subset. This method still achieved an F-score of just under 70%, which is considerably better than the results achieved by Sarafraz and Nenadic [136].

Our system achieved even better results on the GENIA Event (74% F-score) and BioInfer (85% F-score) datasets. This is particularly encouraging, as these corpora contain more complex and varied bio-events than the BioNLP'09 ST corpus.

Our results are also comparable to those obtained by Sanchez-Graillet and Poesio[113], who used a rule-based approach for detecting negated PPIs, and achieved an F-score of 77% with gold standard protein annotations. However, we argue that the identification of negated bio-events in general is a more challenging task, according to the reasons discussed in section 5.1.2.

5.7.2 Selection of Negation Cues

Various lists of negation cues have previously been proposed for the different negation detection tasks. With respect to the task of identifying negated bio-events, the main questions about the nature, role and processing of negation cues are:

Does a "universal" list of negation cues exist?

Our analysis of negated bio-events confirmed that negation cues are ambiguous. Whether a word acts as a negation cue for a bio-event depends on the lexical as well as the contextual polarity of the word. While the lexical polarity of a word remains fixed, its contextual polarity depends on a number of factors including the nature/domain of the text, the annotation perspective, the context and the syntactic structure of the sentence. Therefore, it is hard to compile a universal list of negation cues. However, domain specific lists might be useful. Our experiments provided further evidence for this hypothesis. The c40 and cCore cue lists showed consistently good performance across the three bio-event corpora.

What is the impact of the choice of a negation cue list on the overall system performance?

We designed experiments to measure the impact of the choice of a negation cue list on the overall system performance. We found that a significant variation (ranging between 5% and 8%, depending on the corpus) in the system performance resulted from the cue list used.

Should negation cues be annotated in gold standard corpora?

BioInfer is the only corpus of bio-events containing annotation of negation cues. We compiled a list of negation cues identified in the corpus, and labelled it cBioInfer.

This cue list did not achieve good results when applied to the other two datasets (i.e., GENIA Event and BioNLP'09 ST). However, it outperformed the other cue lists on the BioInfer dataset by a significant margin of 7%. While these results provide further evidence for the domain specific nature of the negation cues, they also highlight the importance of annotating negation cues as well as the polarity of the event. These findings favour the wider argument for the annotation of the lexical cues indicating the information necessary for the correct interpretation of an event.

5.7.3 Feature Engineering and Selection

We have used a novel approach for feature engineering, and have identified an optimum feature set comprising only 17 features, all of which are discrete (14 binary, 1 integer and 2 multi-valued). We have grouped these features into four sets: Semantic, Lexical, Dependency and Command. In comparison to previous work, our feature engineering approach has the following unique aspects:

- use of a combination of semantic, lexical, lexico-semantic and lexico-syntactic features
- use of all available textual fragments associated with the bio-event (including the trigger, participants and attributes of the event)
- use of event hierarchy information (i.e., complexity status)
- use of negation deactivators
- basing the features on the general, rather than specific, dependency relations

An important aspect of our investigation was to evaluate the performance of the individual feature sets as well as their combinations. We were particularly interested in

the comparison of the dependency and the command features, as both have previously been used for the task of identifying negated bio-events. Kilicoglu and Bergler [132] used a rule-based approach based on the dependency relations between the negation cues and the event-triggers, while MacKinlay, Martinez and Baldwin [134] used features derived from the dependency parse of the sentence containing the bio-event. However, Sarafraz and Nenadic [136] used command features to achieve better performance.

The evaluation of the individual feature sets showed that dependency and lexical features achieved results more than twice as high as command features. Similarly, the combination of lexical and dependency features achieves significantly better results than the combination of the lexical and command features. Based on these results, we conclude that, for the task of identifying negated bio-events, dependency features outperform command features by a significant margin. This is consistent with previously reported comparisons between the dependency and constituency features for the tasks of opinion mining [177, 178] and PPI extraction [70].

Aside from the features discussed above, we also experimented with other syntactic and semantic features. We observed that the features based on the POS tags of negation cues, event-triggers, event-themes and event-causes did not improve the performance. Similarly, features based on the semantic types of the event-themes and event-causes did not influence the performance either. This suggests that the polarity status of a bio-event is influenced neither by the semantic types of its participants, nor by the POS tags of text fragments associated with the event.

5.7.4 Algorithm Selection

We designed a series of experiments to evaluate and compare the performance of six learning algorithms with respect to the task of identifying negated bio-events. All of these algorithms, with the exception of Random Forest, had previously been used for different negation detection tasks, with varying degrees of success. Our results showed that, on average, the Random Forest algorithm performs the best; the Decision Trees (C4.5) algorithm scored second by a close margin (1.5%); the Logistic Regression, Instance-Based learning algorithms (IB1) and SVM scored third, fourth and fifth by significant margins of 4%, 6% and 9%, respectively. The Naive Bayes algorithm scored the least by a huge (26%) margin.

Our results are consistent with Caruana and Niculescu-Mizil [179], who conducted a wide ranging study, comparing the performance of ten supervised learning methods. They measured the performance of each method on 11 different binary classification problems, and found that Random Forest outperformed the other algorithms. Our results are also consistent with Goryachev et al. [151], who compared the performance of SVM and Naive Bayes for the task of detecting negations in medical texts. They found that SVM outperformed Naive Bayes by a significant margin (8%). In contrast, Goldin and Chapman [150] compared the performance of Naive Bayes and Decision Trees for the task of identifying negated terms in medical texts. They found that Naïve Bayes outperforms Decision Trees by a small (1%) margin. Similarly, for the task of identifying negation scopes in biomedical research literature, Morante and Daelemans [139] obtained analogous results for Instance-Based learning and SVM. In contrast to these results, we found that Naive Bayes performs significantly worse than Decision Trees, and Instance-Based learning outperforms SVM. This

contrast shows that the different learning algorithms do not perform consistently for different negation detection tasks. This leads us to the following conclusions:

- Despite the apparent similarities, the task of identifying negated bio-events is inherently different from the other negation detection tasks like negated term detection and negation scope detection.
- Since the Random Forest algorithm clearly outperforms the other learning algorithms for the task of identifying negated bio-events, its feasibility for the other negation detection tasks (sections 5.1 and 5.2) should be investigated.

5.7.5 The Effect of Corpus Size

We used all three open access corpora of negated bio-events in our experiments. Table 20 (page 139) shows the statistics for these corpora. The GENIA Event corpus is the largest and contains bio-events of 36 different semantic types. The BioNLP'09 ST corpus contains only 9 types of bio-events, and it is over three times smaller than the GENIA Event corpus. The best results (10-fold cross validation) achieved on the BioNLP'09 ST corpus were 4% less than the best results achieved on the GENIA Event corpus. The BioInfer corpus is the smallest in size (almost 14 times smaller than GENIA Event) and the most complex with 60 different event types. Despite these factors, consistently better results were achieved on BioInfer, irrespective of the cue list used. This suggests that the corpus size does not have a significant effect on overall performance. We further tested this hypothesis by conducting an additional experiment on the GENIA Event corpus. Instead of performing 10-fold cross validation, we trained the classifier using only half of the instances and tested on the

other half. We repeated this experiment ten times with randomly selected training and testing datasets, the average F-score was only slightly (0.5%) less than the F-score achieved by the 10-fold cross validation. Therefore, we conclude that the corpus size is not a significant performance factor. Instead, we believe that the amount of information available about the event, especially the text fragments associated with the event, is more important than the corpus size. The relatively poor performance achieved on the BioNLP'09 ST corpus could also be explained by the fact that both GENIA Event and BioInfer contain more information about the location of the events than BioNLP'09 ST.

5.7.6 Correlation between Event-Type and Polarity

Our analysis of negated bio-events revealed that certain words act as negation cues only in the context of specific types of events. Apart from this, we did not find any evidence of "linguistic correlation" between the semantic type of an event and its polarity. However, we did find some "statistical correlation" between event-type and polarity. For example, in the BioNLP'09 ST corpus, 9% of the *Regulation* events are negated, whereas only 5% of the *Binding* events are negated. Based on this observation, we engineered two semantic features: one based on the event-type and the other on its complexity status (i.e., whether the event is simple or complex). Both of these features scored low gain ratios on all three datasets. However, the addition of these features improved the overall performance by 0.5% to 1%, depending on the dataset.

In order to further investigate the correlation between event-type and polarity, we designed two experiments:

Three-Way Splitting

This experiment was similar to the one reported by Sarafraz and Nenadic [136]. The bio-events in the BioNLP'09 ST dataset were split into three classes according to their level of complexity. The simplest events with single participants, i.e., those of type *Localization*, *Transcription*, *Protein Catabolism*, *Gene Expression* and *Phosphorylation*, were grouped together as *Class-1*. The *binding* events were grouped as *Class-2*. These events have multiple participants, but they only have entities as participants (and not other events). Finally, the most complex events, which allow other events to be participants, were grouped together as *Class-3*. These include both general and specific regulation events, i.e., events of type *Regulation*, *Positive Regulation* and *Negative Regulation*. The Random Forest classifier was trained and tested for each class, separately. The micro averages for precision, recall and F-score were used to measure the overall performance. In comparison to the results achieved without data splitting, the three-way splitting model showed a considerable (21%) improvement in precision. However, the recall dropped significantly (15%), causing an F-score decrease of almost 2%. This is in contrast to Sarafraz and Nenadic [136], who achieved an increase in both recall and precision. In terms of individual classes, *Class-3* and *Class-1* achieved results which were slightly higher and slightly lower than the single-class model, respectively. However, *Class-2* scored significantly (29%) worse. We experimented with various algorithms and cue-lists, but we were not able to improve the performance for *Class-2* by more than 2%.

The above results can be explained by considering the uneven distribution of events within the three classes. For example, in the BioNLP'09 ST corpus, 56% of events belong to *Class-3*, 33% to *Class-1*, and only 11% to *Class-2*. Similarly, the distribu-

tion of negated events within each class also varies: 5% for *Class-1*, 4% for *Class-2* and 9% for *Class-3*. Therefore, better results would be expected for classes with higher numbers of training examples (e.g., *Class-3*) and vice-versa (e.g., *Class-1*).

Two-Way Splitting

In this experiment, we split the bio-events according to their complexity status, i.e., *simple* or *complex*. We performed the two-way splitting on the BioNLP'09 ST data, then trained and tested our Random Forest classifier separately for each class. The results were even worse than the three-way splitting model, and an overall (micro-averaged) performance loss of 5% was observed. In order to test the concept further, we repeated the two-way splitting experiment with the GENIA Event corpus. Again, we observed a significant (4%) decrease in performance. In terms of individual classes, the *complex* class performed better than the *simple* class. We further experimented with various algorithms and cue-lists, but we were not able to improve the performance on the *simple* class by more than 1%. We also observed that over 10% of *complex* events are negated, where only 4% of *simple* events are negated. Therefore, a *complex* event is 2.5 times more likely to be negated than a *simple* event.

These experiments show that splitting the datasets according to the event-type does not improve the overall system performance. The classification performance improves for certain classes of bio-events (e.g., *complex* event and *regulation* events), and deteriorates for certain other classes (e.g., *binding* and *simple* events). This variation in performance is mainly due to an uneven distribution of negated bio-events across these classes.

5.8 Conclusion

We have conducted a detailed analysis of the problem of identifying negated bio-events given gold standard event annotations. We examined the manifestations of negation in the three open access corpora of negated bio-events (i.e., GENIA Event, BioInfer and BioNLP'09 ST), and proposed a typology of negated bio-events based on the lexico-semantic mechanisms affecting the polarity of an event. Our analysis showed that a significant proportion (37%) of negated bio-events cannot be detected by considering the event-trigger alone. It also revealed that identification of negated bio-events is a complex task that requires a deeper level of analysis than that required for tasks such as negated term detection and negation scope detection. Following this examination, we identified the three key aspects of a machine learning based solution to the problem of negated bio-event detection. These are: the compilation of a negation cue list, the design and selection of suitable features and the choice of a machine learning algorithm. In order to analyse these aspects, we conducted a series of experiments on the three bio-event corpora. The results confirmed that each one of these aspects can have a significant impact on the overall system performance. Our analysis showed that the ability of a word/phrase to act as a negation cue depends not only on the context and domain of text, but also on the annotation/information perspective (e.g., linguistic vs. biological perspective). Therefore, there is a need for domain specific lists of negation cues. We compiled two such lists (c40 and cCore), both of which performed consistently in all experiments. In terms of feature selection, our results showed that lexical and dependency features are most important, while command and semantic features are less significant. Nonetheless, the best results were achieved by a combination of all four types of features.

We also discovered that, for this task, the Random Forest algorithm consistently outperforms the other learning algorithms. Combining the best solutions for each of the above aspects, we created a novel framework for the identification of negated bio-events. We evaluated our system on the three open access corpora of negated bio-events mentioned above. Our results on the BioNLP'09 ST corpus were significantly higher than the previously reported best results. We achieved even better results on the GENIA Event and BioInfer corpora, both of which contain more varied and complex events.

As mentioned earlier, our system assumes that event annotation has already been performed. However, this system can be integrated with a state-of-the-art event extraction system, e.g., EventMine (section 2.2.3). The resulting system will be able to extract bio-events of the specified polarity from plain text documents, and it will serve as the foundation for a more elaborate system for detecting textual contradictions.

Chapter 6: Manner of Bio-events

6.1 Introduction

In this chapter, we describe the design and evaluation of a machine learning system that can automate the assignment of a further *dimension* of meta-knowledge to bio-events, i.e. *Manner*. This is the most domain-specific dimension of our scheme, which encodes the rate, level, strength or intensity of the event (in biological terms). The detection of manner information can be useful for several tasks, e.g., in comparing results obtained by different authors, or to help to detect possible contradictions or inconsistencies in the results reported in different papers. The identification of such information is considered to be highly important for the correct interpretation of biomedical events [180]. To our knowledge, our system is the first that is able to automatically identify and classify information about manner in biomedical text, through the assignment of three possible values to events, i.e., *High, Low* and *Neutral*, with the latter being the default value. Given that non-default manner values are assigned to around 5% of events in the GENIA-MK corpus [9], a majority class baseline system would achieve an accuracy of 95%. Through the employment of a combination of several different feature types, i.e., syntactic, semantic, lexical, lexico-semantic and lexico-syntactic, our system is able to perform considerably better than the baseline, with an overall accuracy of 99.4% and micro averaged F-scores of 98.3%.

6.1.1 Manner of Bio-Events

The term "manner" could correspond to any information about *how* an event occurs, and so is not in itself domain-specific. Indeed, manner is annotated as a general adjunct-like argument type in the PropBank corpus, [181], which provides a semantic annotation of general language verbs that appear in the Penn Treebank [182]. However, since adjuncts are considered to be general phrases that are not closely associated with any particular verb, they are not normally specified in semantic frame resources that are developed for general language.

In contrast, manner is considered to be highly important for the correct interpretation of biomedical relations and events [180]. Accordingly, in the GREC corpus [32], *Manner* was annotated as one of 13 fixed semantic roles that can characterise the semantic arguments of verbs and nominalisations in biomedical texts. The annotations were extracted as semantic frames and linked with syntactic frames in the BioLexicon [183], thus allowing the identification of verbs that are particularly likely to specify manner information in biomedical texts.

In the GREC corpus and the BioLexicon, the characterisation of manner arguments can be quite wide-ranging. They can correspond to the intensity of an event. However, they can also correspond to a process or method that is employed by the agent to bring about the event (normally a noun phrase following the preposition *by*), an adverb relating to a process that describes how the event is carried out, information about the direction of an event, etc.

As has been explained earlier, each dimension of event meta-knowledge comprises a fixed set of values, e.g., there are 2 possible values for *Polarity*, and 3 for *Certainty Level*. Thus, while the BioLexicon can help to identify diverse phrases that are relat-

ed to the manner of an event, the *Manner* dimension in our meta-knowledge scheme aims to provide a useful *classification* of events according to the type of manner that they express. Given the wide range of information that can come under the general heading of manner, our meta-knowledge scheme focusses on a restricted view of the manner of biological processes, which lends itself to a reasonably straightforward division into a set of distinct categories, and which are feasible to recognise automatically.

We took as our starting point the relatively narrow definition of manner proposed in [113] for a specific type of bio-event, i.e., Protein-Protein Interactions (PPI). According to them, manner may reveal levels of interaction or certainty of the reported interaction, and is indicated by *manner cues* (adjectives or adverbs) that affect the PPI *trigger* (the word or phrase indicating the presence of a PPI). Based on our analysis of bio-events, our definition of manner is a slightly modified version of the one provided in [113]. Firstly, we did not include aspects of certainty, since we treat *Certainty Level* as a separate meta-knowledge dimension. Secondly, we extended the other part of the definition slightly, to cover information concerned with the rate, strength or intensity of the event, as well as the level. This expanded interpretation is needed, given that our meta-knowledge annotation scheme is intended to be applicable to a wider range of events than only PPIs, whose varying semantics mean that expressions of manner can have subtly different interpretations according to the type of event they modify. Based on a manual examination of over 100 abstracts in the GENIA Event corpus, we found that events can normally be ascribed to one of the following three categories of manner (see section 3.3.5 for further details and examples of these categories):

- **High:** The event has explicit indication of higher than default rate, level, strength or intensity. Cue expressions are typically adjectives or adverbs such as *high*, *strongly*, *rapidly*, *potent*, etc.
- **Low:** The event expresses lower than default rate, level, strength or intensity. Cue expressions are typically adjectives and adverbs such as *slightly*, *partially*, *small*, etc.
- **Neutral:** The default category, for events with no explicit indication of either *High* or *Low* manner. In rare cases, *Neutral* manner is explicitly indicated, using cue words such as *normal* or *medium*, etc.

When combined with polarity, annotation of event manner can help to capture subtle variations between the interpretations of different events. That is to say, a distinction can be made between "low interaction" and "no interaction". Historically, certain cues of *Low* manner (like *low*, *little*, *small*, etc.) have been treated as negation indicators. In the field of sentiment analysis, these cues have been considered as a special class of negative polarity indicators, which have been referred to as both *diminishers* [156] and *negative polarity shifters* [157]. The same types of cues have been treated as negation triggers in the field of biomedical text mining [29, 31]. However, in the context of bio-events, there is a clear and important distinction between a *Low* manner event and a negated (i.e., non-existent) event. This view has been confirmed by biologists who were consulted and involved in the creation of the GENIA-MK corpus.

6.1.2 Annotation of Manner in the Enriched GENIA Event Corpus

As discussed in chapter 4, our analysis of the meta-knowledge annotations in the GENIA-MK corpus revealed that 1,392 events (4%) are expressed with *High* manner, 323 events (1%) are expressed with *Low* manner, and the remaining 35,143 events (95%) were found to be of *Neutral* manner. Amongst events with an explicit indication of manner, *High* manner marking is much more common, accounting for 81% of cases. However, the significance of identifying instances of *Low* manner cannot be overlooked, since, as described above, it can help to distinguish between truly negative events and those that occur at a low level or with low intensity. Interestingly, the overall frequency of events expressed with a non-default manner is only 1% less than the frequency of negated events [9]. While negation detection has received significant attention in the literature [184], manner identification in biomedical text remains an understudied area of research.

6.2 Automated Identification of Event Manner

Since manner is considered an important part of biomedical event descriptions, it follows that training a system to classify events according to the type of manner they express is an important task. To our knowledge, the automatic classification of manner-related information has not previously been attempted in biomedical text, either at the level of events or for larger units of text.

6.2.1 Analysis of Manner Cues

The textual context of an event and the syntactic structure of the sentence in which the event is contained can both play important roles in determining the most appropriate manner value to assign to an event. Accordingly, these are both taken into ac-

count by the set of features used by our classifier, as explained in the next section. However, the single most important factor is the presence of an explicit cue expression in a sentence. Thus, we carried out a detailed analysis of the manner cues identified in the GENIA-MK corpus. Some of the key findings are as follows:

6.2.1.1 Cue Frequency

While a total of 273 *High* and 103 *Low* manner cues have been identified, most of these cues (72%) appear just once or twice, and only a handful (9%) appear 10 or more times. Moreover, this small set of the most frequent cues occur in the textual context of the majority (61%) of events that are expressed with a non-default manner. These statistics demonstrate that although a relatively small set of cues accounts for a majority of *High/Low* events, much larger cue sets need to be considered in order to achieve optimum results for automated manner identification.

6.2.1.2 Cue Variation

While most cues for non-default manner consist of particular words and phrases, others constitute patterns, in which different numerical values may be substituted. An example is the expression *n-fold,* in which *n* represents a number. This expression accounts for 111 (over 8%) of the *High* events. However, a particular challenge lies in the fact that the exact form of expression can vary. Indeed, in the GENIA-MK corpus, 13 different variants of this numerical expression have been annotated as *High* cues. Some examples include *2-fold*, 4-*6 fold*, *5- to 7-fold*, etc. Moreover, four non-numeric variants (*two-fold*, *threefold*, *two to threefold* and *two-three fold*) have also been annotated as *High* cues. These non-numeric variants account for a further

14 *High* events. Similarly, several variants of the numeric expression *n%* have also been annotated as both *High* and *Low* manner cues.

6.2.1.3 Cue Ambiguity

The presence of a *High/Low* cue in a sentence is not sufficient to assign a *High/Low* value to all events in the sentence. While a sentence contains, on average, four bio-events, the majority of manner cues affect only one event in the sentence. Therefore, the syntactic structure of the sentence needs to be considered to determine which, if any, events are being affected by the cue. The semantic context also plays an important role in determining the identity of some cue expressions. For example, depending on the context, numerical expressions (like *n-fold* and *n%*) may indicate a *High* manner, a *Low* manner or neither.

6.2.1.4 Combined Event-Triggers / Manner Cues

Whilst most manner cues are independent of event type, certain words can act simultaneously as both event-triggers (which denote the type of the event) and manner cues. For example, the word *overexpression* is an event-trigger that introduces an event of type *Gene Expression*. Furthermore, the word tells us that the event occurred with *High* manner.

6.2.1.5 Effect of Negation

An expression of negation inverts the polarity of a manner cue. For example, the word *significant* acts as a *High* cue, but its negated form (*no/not significant*) is a *Low* cue.

6.2.2 Classifier Design

In this section, we explain the various different types of features that are used by our classifier, together with an explanation of the learning algorithm that was employed.

6.2.2.1 Features

We used a combination of syntactic, semantic, lexical, lexico-semantic and lexico-syntactic features. The Enju parser [173] was used to obtain the lexical and syntactic information required to construct these features. We also compiled master cue lists for the *High* and *Low* categories by extracting all *High/Low* cues identified in the GENIA-MK corpus. These cue lists were also used in the generation of features. A brief explanation of each feature set is as follows:

Syntactic Features

Syntactic features include the POS of the event-trigger, event-participants and the *High/Low* cues found in the sentence.

Semantic Features

These are constructed from the semantic information that is annotated for the bio-event. They include the semantic type of the bio-event (e.g., *Gene Expression*, *Positive Regulation*, etc.), the semantic type of each participant (e.g., *lipid*, *DNA molecule*, etc.) and the role of each participant (e.g., *theme* and *cause*, etc.). We have also used a *complexity* feature, which indicates whether a bio-event is simple or complex. The latter value means that the event has one or more participants which are bio-events themselves.

Lexical Features

These include the presence of a *High/Low* cue in the sentence, the cue itself, the presence of a negation indicator and its relative position with respect to the *High/Low* cue, etc. We used regular expressions to identify numeric cues, such as *n-fold* and *n%.*

Lexico-Semantic Features

These are constructed using a combination of the "textual" bio-event information and information from the sentence containing the bio-event. The textual bio-event information includes the text fragment indicating the occurrence of the bio-event (i.e., the event-trigger), the text fragments identifying the event participants and the text fragments indicating any event attributes like location, etc. The features used include the surface distances between the *High*/*Low* cue and the event-trigger, participants and event-location, whether the *High/Low* cue is part of the event-trigger, and whether the *High/Low* cue precedes or follows the event-trigger, etc.

Dependency (Lexico-Syntactic) Features

These are constructed using the textual bio-event information and the dependency relations in the sentence identified by the Enju parser. These features include the presence of direct and indirect dependency relations between the *High/Low* cue present in the sentence and the event-trigger and/or event-location, the types of the dependencies and the lengths of the dependency paths.

Constituency (Lexico-Syntactic) Features

These are based around the *command* [158] and *scope* relations, which are derived from the constituency parse tree. We used several command features including the

existence of S-, VP- and NP-command relations between the *High/Low* cue and the event-trigger, and/or event-participants. The scope features consider whether the event-trigger falls under the syntactic scope of the *High/Low* cue.

6.2.2.2 Learning Algorithm

Given its superior performance in the task of identifying event polarity (chapter 5), we decided to build the classifier using the Random Forest [167] algorithm. As previously explained, this algorithm develops an ensemble/forest of Decision Trees from randomly sampled subspaces of the input features. Once the forest has been created, new objects are classified by first obtaining individual classifications from each tree and then using a majority vote to attain the final classification. The Random Forest algorithm has been successfully used for various text mining and bioinformatics tasks [168, 169]. We used the WEKA [174] implementation of the Random Forest algorithm, which is based on [167]. Our optimisation settings included: (1) setting the number of trees in the forest to 10, (2) setting the number of features used to build individual trees to log(N+1), where N is the total number of features, (3) setting no restrictions on the depth of individual trees.

6.3 Results and Discussion

We conducted a series of experiments using different cue lists and feature combinations. All results were 10-fold cross validated. The best results, as shown in Table 28, were achieved using all feature sets (mentioned in section 6.2), the 50 most frequent *High* cues and the 25 most frequent *Low* cues.

Although reasonable results (71% F-score) were achieved for the *Low* category, the results for the *High* category were significantly better. This is partly because the

number of training examples available for the *High* category is 4 times higher than those available for the *Low* category. Moreover, the *Low* cues are more diverse and scattered than the *High* cues. The best results were achieved for the *Neutral* category. However, this is to be expected, given that the vast majority of training examples belong to this category. In order to evaluate the overall classifier performance, we calculated the macro and micro averages. The micro averaged results were significantly higher than the macro averaged results. This is because the best classified category (*Neutral*) is also the most abundant by a significant margin.

Category	Precision	Recall	F-Score
High	85.1%	77.7%	81.2%
Low	78.7%	65.4%	71.4%
Neutral	99.1%	99.4%	99.2%
Macro Avg	87.6%	80.8%	83.9%
Micro Avg	98.4%	98.3%	98.3%

Table 28. Classification Results (10-fold CV)

As mentioned above, since 95% of all events belong to the *Neutral* category, a classifier which assigns the *Neutral* category to all instances will achieve an accuracy of 95%. Therefore, this figure provides a natural baseline for measuring the overall accuracy of the classification system. Our classification system achieved an overall accuracy of 99.4%, which is significantly higher than the baseline.

For the *High* category, the recall is 7% lower than precision. This difference is almost double (13%) for the *Low* category. An error analysis revealed that, for both categories, the main factor contributing towards reduced recall was the inability of

the system to identify the *High/Low* cues present in the sentence. As mentioned above, cues are mainly identified via *High/Low* cue lists. Given the ambiguous nature of *High/Low* cues, the size of these lists introduces a precision-recall trade-off, i.e., larger cue lists improve recall at the expense of precision. Thus, the optimum results (as shown in Table 1) were achieved using cut-down versions of the master cue lists. The use of shorter cue lists (i.e., the 50 most frequent *High* cues and the 25 most frequent *Low* cues) enhanced the classification performance (F-score) by 5% for the *High* category and by 7% for the *Low* category. However, it imposed implicit upper-limits of 91% and 79% on the recall for the *High* and *Low* categories, respectively.

A significant proportion (23%) of misclassified events belonged to sentences with complex syntactic structures, e.g., where the event-trigger and the *High/Low* cue belonged to different clauses. These misclassifications can be partly attributed to parsing limitations, especially in terms of identifying complex dependency relations.

6.4 Conclusion

We have analysed the problem of the identification of manner in bio-events and have presented a machine learning based solution to this problem. We have shown that the manner of bio-events can be automatically identified with a high degree of accuracy. Our classification system achieves an overall accuracy of over 99% and macro and micro averaged F-scores of 84% and 98% respectively. Given the level of accuracy achieved by our system, it can be applied to enrich other bio-event corpora with manner information automatically. The manner identification system can be integrated with an event extraction system (section 2.2.2). The resulting system will be able to extract bio-events with the specified manner type from textual sources.

Chapter 7: Knowledge Source of Bio-events

7.1 Introduction

In recent years, several annotation schemes, e.g., [102, 103, 110, 185] have been developed to identify and classify textual zones (i.e. continuous spans of text, such as sentences and clauses) in scientific papers, according to their rhetorical status or general information content. In most cases, these corpora have subsequently been used as a basis for training systems to recognize this information automatically, e.g. [111, 186, 187]. Common to all of these systems is the ability to identify information about knowledge source. That is, whether the text zone refers to new work being described in the paper, or refers to work that has already been described elsewhere. Such systems can be instrumental in helping users to search for text zones that contain new experimental knowledge. The identification of such information is important for several tasks in which biologists have to search and review the literature. One such example is the maintenance of models of biological processes, such as pathways [29]. As new reactions or new evidence for reactions become available in the literature, these should be added to the corresponding pathway(s). Another area where this information is useful is in the curation of biomedical databases. One of the tasks involved in keeping such databases up to date is to search for new evidence for a particular interaction (e.g., gene regulation) within the literature [92].

In the types of task outlined above, the biologist is likely to be looking for specific types of biological processes or reactions, and specific types of information about them, e.g., what caused the reaction to occur, where the reaction took place, etc. Text

zone classification systems cannot help with this kind of task. However, event extraction systems can be extremely useful in this situation.

As explained in section 2.3, event extraction systems can facilitate the development of sophisticated semantic search systems, e.g., [67], which allow researchers to perform structured searches over events extracted from a large body of text [188]. Although search constraints can typically be specified in terms of event type (i.e., the process or reaction of interest) and/or the types of named entities participating in the event, the ability to specify knowledge source as a constraint is not available. Bio-events are typically contained within a single sentence, and text zone identification systems would normally be able to determine knowledge source at the sentence level. However, events are not the same as text zones. Whilst text zones constitute continuous spans of text, events usually consist of several discontinuous text spans, consisting of components identifying the event [29]. There are also (usually) several events contained within a single sentence. This means that just because a sentence or clause may be identifiable as having a particular knowledge source, it does not follow that the events contained within that text zone will all have the same knowledge source; each event may have its own interpretation, and determining which events are affected by particular textual cues can be complex.

In the remainder of this chapter, we describe our work on the analysis and automated identification of knowledge source information about bio-events, using the GENIA-MK (abstracts) and FP-MK (full papers) corpora for training and testing. In both corpora, each event is ascribed one of two knowledge source values, i.e., *Current,* for events relating to work described in the current paper (default value), or *Other,* for events relating to work originally described elsewhere. Although our previous

analysis [18] revealed that there are significant differences in the distributions of the different knowledge source values in abstracts and full papers, and that the textual means of denoting *Other* events also varies between abstracts and full papers, our system is able to perform to an almost identical level of accuracy on both text types, i.e., 99.6% and 99.4%, for abstracts and full papers, respectively.

7.1.1 Knowledge Source

As mentioned above, information about knowledge source is an integral part of a number of schemes for annotating text zones and their functions. The argumentative zoning (AZ) scheme, first introduced in [103], distinguishes sentences that mention OWN work presented in the current paper and OTHER specific work presented in another paper. Later extensions based on this scheme [102, 189] recognised that different types of information about OWN work can usefully be distinguished, such as OWN_METHD (methods) and OWN_RES (results) or OWN_CONC (conclusions). Multi-dimensional schemes allow several pieces of information to be associated with a given text span, and thus provide more flexibility regarding the types of information that can be encoded. Several such schemes encode information about knowledge source as a separate dimension, e.g., the scheme of [187] includes a novelty attribute (*New* or *Old*) that is distinct from their knowledge type attribute (*Background, Method, Conclusion*, etc.). The scheme of [110] identified five dimensions of information that could reliably be identified about text fragments (mostly clauses or sentences). Their *evidence* dimension includes information about the source of knowledge expressed in the text fragment. It has four possible values, which have similarities with some of the evidence codes used during the annotation for the Gene Ontology [30]. These values are: *E0*: no indication of evidence; *E1*: mention of evi-

dence with no explicit reference; *E2*: explicit reference is made to other papers to support the assertion; *E3*: experimental evidence is provided directly in the text.

In our model of event interpretation [8], information about the knowledge source of the event is encoded using the *Knowledge Source* dimension, which has two possible values. The **Other** value is assigned when the event can be attributed to a previous study. This value is normally determined through the presence of explicit cues, e.g., *previously, recent studies*, etc., or cited papers, in the vicinity of the event. The **Current** value is assigned when the event makes an assertion that can be attributed to the current study. This is the default category, and is assigned in the absence of explicit lexical or contextual cues, although explicit cues such as *the present study* may be encountered. Further details and examples can be found in section 3.3.5.

7.1.2 Annotation of Knowledge Source in GENIA-MK and FP-MK Corpora

As discussed in chapter 4, the GENIA-MK corpus consists of 1,000 MEDLINE abstracts, containing 36,858 events, each of which has been annotated according to our meta-knowledge scheme [18]. In this corpus, only 1.5% of all events are assigned a *Source* value of *Other*. This is not surprising: abstracts are meant to provide a summary of the work carried out in a given paper and (given the very limited space) there is little opportunity to discuss previous work. Indeed, the use of citations is often prohibited in abstracts.

The FP-MK corpus consists of 4 full papers, in which 1,710 events have been annotated according to the same meta-knowledge scheme. In contrast to the GENIA-MK corpus, nearly 20% of all events in the FP-MK corpus belong to the *Other* category.

Our analysis [18] showed that by far the highest concentration of *Other* events is in the *Background* sections of the papers, where over 40% of the events are attributed to other sources. This is expected, since it is normally in the *Background* section where one encounters the highest concentration of descriptions of previous work. The *Discussion* sections of the papers also have a high (over 25%) concentration of *Other* events, since it is common to compare and contrast the outcomes of the current work with those of previous related studies as part of the discussion. The frequency of *Other* events in the remaining sections is considerably lower. For example, in the *Results* sections of the papers, less than 7% of events are annotated as *Other*. Further details and a comparison of knowledge source annotation in the GENIA-MK and GENIA-FP corpora can be found in chapter 4.

7.2 Analysis of Other Events

7.2.1 Cue Frequency

Table 29 shows the most commonly annotated cue expressions for *Source=Other* in the GENIA-MK (abstracts) and FP-MK (full papers) corpora respectively. For abstracts, cue expressions contain the adverbs *previously* or *recently*, or their adjectival equivalents. The phrases *have been* and *has been* have also been annotated as cues with reasonably high frequency, the reason being that the use of the passive voice with the present perfect tense (e.g., *has been studied)* is a common means to indicate that an event has previously been completed (e.g., in a previous study), but yet has relevance to the current study.

In contrast to abstracts, the vast majority of cue expressions in full papers correspond to citations. However, similarly to abstracts, the use of past perfect tense is

also quite common. Other explicit markers (such as *previously* and *recently*) constitute less than 10% of the cue expressions.

GENIA-MK (abstracts)			FP-MK (full papers)		
Cue	**Freq**	**%**	**Cue**	**Freq**	**%**
previously	118	21.7%	Citation	267	78.3%
has been	89	16.3%	has been	41	12.0%
recently	67	12.3%	previously	6	1.8%
have been	39	7.2%	recently	6	1.8%
previous	38	7.0%	latter example	4	1.2%
recent	32	5.9%	studies have shown	4	1.2%
earlier	6	1.1%	we and others	4	1.2%

Table 29. Most frequently annotated *Other* cues in GENIA-MK and FP-MK corpora

7.2.2 Cue Ambiguity

Similarly to the other meta-knowledge cues, the presence of an *Other* cue in a sentence is not in itself sufficient evidence for assigning the knowledge source value of *Other* to all events in the sentence. While a sentence contains, on average, 4 bio-events, the majority of *Other* cues affect only one event in the sentence, i.e., the knowledge source value for the remaining events in the sentence is *Current* (not *Other*). Therefore, it is highly important that the syntactic/semantic structure of the sentence is considered, in order to determine which, if any, events are being affected by the cue. For example, the existence/type of dependency/constituency relations between the event participants and any *Other* cue(s) present in the sentence can be considered.

Furthermore, some of the *Other* cues (e.g., the tense of the sentence) are inherently ambiguous, and only indicate an *Other* event in certain contexts. For example, the cue expression *has/have been* is a significant cue for *Other* events – it accounts for over 23% of all *Other* events in abstracts and 12% of all *Other* events in full papers. However, an analysis of events from the sentences containing the phrase *has/have been* in the GENIA-MK corpus reveals that only 8% of these events are of type *Other*. This proportion is even lower (7%) for full papers.

7.2.3 Event Complexity

We examined the distribution of events assigned the value *Source=Other* amongst **simple** and **complex** events. As explained earlier, by simple event, we mean an event whose participants are all entities, whilst a complex event is one with at least one participant which is itself an event. In abstracts, 67% of *Other* events are complex. Conversely, 2.26% of complex events are of type *Other*, while only 0.88% of simple events are of type *Other*. This means that an arbitrary complex event is 2.6 times more likely than an arbitrary simple event to have knowledge source value of *Other*.

In full papers, an even greater proportion of *Other* events (i.e., 72%) is complex. A total of 3.32% of complex events are of type *Other*, while only 0.73% of simple events belong to this type. Therefore, in full papers, an arbitrary complex event is 4.5 times more likely than an arbitrary simple event to have knowledge source value of *Other.*

7.2.4 Relative Position within Text

In abstracts, 74% of *Other* events appear in the 2nd, 3rd or 4th sentence. Furthermore, over 80% of the *Other* events appear in the first half of the abstract.

In full papers, the section to which the sentence containing the event belongs is more significant than the relative position of the sentence within the paper or even within a section. For example, over 60% of all *Other* events found in full papers occur in the *Background* section.

7.3 Classifier Design

Based on the analysis of *Other* events, we engineered 7 feature sets. We used the Enju parser [173] to obtain the lexical and syntactic information required to construct these features. A brief explanation of each feature set is as follows:

- **Syntactic features** include the tense of the sentence, the POS tag of the *event-trigger*, and the POS tag(s) of *Other* cue(s) found in the sentence.
- **Semantic features** include the type of the bio-event and the type and role of each participant.
- **Lexical features** include the presence of an *Other* cue in the sentence and the cue itself. We used a combination of cue lists extracted from the two corpora and regular expressions to identify *Other* cues.
- **Lexico-semantic features** include the surface distances between the *Other* cue and the event components (*event-trigger, event-participants* and *event-location*), whether the *Other* cue precedes or follows the *event-trigger*, etc.
- **Dependency (lexico-syntactic) features** are based around the presence of direct and indirect dependency relations between the *Other* cue present in the sentence

and the *event-trigger*, and the length of these dependency paths.

- **Constituency (lexico-syntactic) features** are based around the *command* [158] and *scope* relations, which are derived from the constituency parse tree. The command features consider the existence of S-, VP- and NP-command relations between the *Other* cue and the *event-trigger*. The scope features consider whether the *event-trigger* falls under the syntactic scope of the *Other* cue, i.e., whether (on the syntactic parse tree) the node representing the *event-trigger* is a descendant of the node representing the *Other* cue.
- **Positional features** include the section in which the sentence containing the event appears (for abstracts all events have the same value and this feature becomes redundant), and the relative position of the sentence containing the event within the entire text and within the section.

Given its consistent performance in identification of other meta-knowledge dimensions, we used the WEKA [174] implementation of Random Forest [167] algorithm, which is based on [167]. We used the same optimisation settings as for manner detection: (1) setting the number of trees in the forest to 10, (2) setting the number of features used to build individual trees to log(N+1), where N is the total number of features, (3) setting no restrictions on the depth of individual trees.

7.4 Results and Discussion

We conducted a series of experiments using different cue lists and feature set combinations. All results were 10-fold cross validated. The best results for abstracts and full papers are shown in Table 30. In both cases, the best results were achieved by using the 7 most frequent cues and all feature sets.

Category	GENIA-MK (abstracts)			FP-MK (full papers)		
	Precision	**Recall**	**F-score**	**Precision**	**Recall**	**F-score**
Current	99.6%	99.8%	99.7%	99.5%	99.2%	99.3%
Other	83.3%	70.8%	75.6%	81.3%	70.1%	75.3%
Macro-Average	91.5%	85.3%	88.1%	90.4%	84.7%	87.3%
Micro-Average	99.4%	99.4%	99.4%	95.9%	93.4%	94.6%

Table 30. Best results for GENIA-MK and FP-MK

7.4.1 Abstracts

In abstracts, only 2% of all events are of type *Other*; therefore, the baseline accuracy (through majority-class allocation) is 98%. Our system achieves an overall accuracy of 99.6%, which is considerably higher than this baseline. Recall for the *Other* category is significantly lower than the precision (over 10%). This is mainly due to the difficulty in identifying and disambiguating *Other* cues. The overall system precision and recall are both 99.4%.

7.4.2 Full Papers

The proportion of *Other* events in full papers is almost 10 times greater than in abstracts, with just under 20% of all events belonging to the *Other* category. The baseline classification accuracy for full papers is thus 80%. Therefore, statistically, identification of knowledge source in full papers is a harder task than in abstracts. However, our system achieves a very high overall accuracy of 99.4%. The main difference between the *Other* events in abstracts and full papers is the occurrence of explicit citations as clues. Since our system also includes citation related features, it is able to perform equally well on both corpora.

Similarly to the results for abstracts, precision for full papers is significantly higher than recall. Again, this is mainly due to the difficulty in identifying/disambiguating *Other* clues. This is also reflected in overall system performance as well, where precision is 2.5% higher than recall.

7.4.3 Discussion

These results are the first that concern the detection of knowledge source at the event level. However, some comparisons can be drawn with similar previous work at the clause, sentence, and zone level. The zone classification system of [186] achieved a precision/recall of 51%/30% for their OTHER category and a precision/recall of 85%/86% for the OWN category at the text zone level. [190] achieved an overall F-score of 70% for automatic zone classification, including BACKGROUND and OWN zones. The clause classification system reported by [111] performed with F-scores of 89%, 57%, 94% and 91% for the E0, E1, E2, and E3 classes respectively. [187], whose classification is performed at the sentence level, achieved an F-score of 64% for their BACKGROUND class; however, they did not try to identify the novelty attributes separately. Although we identify knowledge source at the event level, which is more challenging than similar tasks at the clause/sentence/zone level, our results are significantly higher. This is partly because we have cast the problem as a binary classification rather than a multi-category classification.

In our system, the most common reason for misclassification was the inability of the system to identify *Other* cues. This accounted for over 52% of the misclassified events. A significant proportion (23%) of misclassified events belonged to sentences with complex syntactic structures, e.g., where the *event-trigger* and the *Other* cue

belonged to different clauses. These misclassifications can be partly attributed to parsing limitations, especially in terms of identifying complex dependency relations.

7.5 Conclusion

The isolation of new experimental knowledge in large volumes of text is important for several tasks undertaken by biologists. Although the ability to search for events of interest can significantly reduce the biologist's workload in finding relevant information, even more time could be saved if it was possible to identify only events pertaining to reliable new experimental knowledge. This goal can be achieved through the automatic recognition of event meta-knowledge. One of the most crucial aspects of identifying new experimental knowledge is to determine the knowledge source of the event. We analysed the event-level knowledge source annotations in the GENIA-MK corpus (abstracts) and the FP-MK corpus (full papers). This analysis was used to inform the process of designing a system to recognize this information automatically. We have shown that the knowledge source of events can be recognized to a high degree of accuracy. In abstracts, the overall accuracy is 99.6%, with macro and micro averaged F-scores of 88.1% and 99.4%, respectively. The baseline accuracy for abstracts is already extremely high (98%), given that there are few events in abstracts that refer to previous work. However, a more significant result is that the performance of the classifier on full papers is almost as high as for abstracts, even though the baseline accuracy for full papers (80%) is considerably lower than for abstracts. On full papers, the classifier performs with an overall accuracy of 99.4%, with macro and micro averaged F-scores of 87.3% and 94.6%, respectively. These results provide encouraging evidence that the knowledge source of biomedical events can be predicted very reliably, regardless of text type.

Chapter 8: Meta-knowledge based Discourse Analysis

Annotating biomedical text with discourse-level information is a well-studied topic. Several research efforts have annotated textual zones (normally sentences or clauses) with information about their rhetorical status, whilst other efforts have linked and classified sets of text spans according to the type of discourse relation holding between them. We have investigated a new approach to discourse analysis, which involves annotating both rhetorical intent and other types of information (such as certainty level and knowledge source) at the level of bio-events. In this chapter, we report on the examination and comparison of transitions and patterns of event meta-knowledge values that occur in both abstracts and full papers. Our analysis highlights a number of specific characteristics of event-level discourse patterns, as well as several noticeable differences between the types of patterns that occur in abstracts and full papers.

8.1 Introduction

The identification of information about the structure of scientific texts has been studied from several perspectives. One line of previous research has been to classify textual zones (e.g., sentences or clauses) according to their function in the discourse, such as background knowledge, hypotheses, experimental observations, analyses, conclusions, etc. The automatic identification of such information can help in tasks such as isolating new knowledge claims in a research paper [191]. Several annotation schemes, e.g., Teufel et al.[186]; Mizuta et al. [102]; Wilbur et al.[110]; de

Waard & Pander Maat [192]; Liakata et al. [185], have been developed to classify textual zones (i.e., continuous spans of text, such as sentences and clauses) according to their rhetorical status or general information content. Sentences and clauses in text are usually not understood in isolation, but rather in relation to other sentences and clauses [193]. Therefore, for certain tasks, such as automatic summarisation, it is important to gain a fuller understanding of how the different types of information conveyed in the text are arranged to form a coherent discourse. This can involve analysing the arrangement or progression of different types of textual zones within a document. For example, Swales [194] defined a model that describes the structure of the introductions to scientific articles, consisting of 3 different fixed moves, with a total of 11 possible steps, each of which normally corresponds to a sentence or clause. Teufel [186] examined patterns of argumentative zones that occur in scientific abstracts.

A further approach to discourse analysis has been to identify sentences and clauses that are linked together, in terms of discourse, and to determine how these links/relations should be characterised. Several efforts to produce annotated corpora or systems to detect discourse structure automatically have been based around the Penn TreeBank corpus of open domain news articles [182]. Carlson et al. [195] enriched the Penn TreeBank corpus with hierarchically structured discourse trees, based on Rhetorical Structure Theory (RST) [196], which uses 78 different discourse relations types, falling under categories such as *Background, Cause* and *Comparison*. Marcu & Echihabi [193] created a system to predict certain classes of discourse relations automatically. The Penn Discourse TreeBank (PDTB) [197] added discourse relations to the Penn TreeBank, both implicit and explicit, that hold between

pairs of text spans. The relations are assigned senses from a hierarchical scheme, such as *Cause* and *Condition*. The Biomedical Discourse Relation Bank (BioDRB) [198] annotates the same types of relations as the PDTB in biomedical research articles. Discourse analysis approaches based on dependency relations have not been restricted to the English language only. Dependency tree banks have also been created for several languages including Arabic [199], Chinese [200], and Czech [201].

All of the studies above considered sentences or clauses as the units of annotation. In contrast, our work is concerned with discourse information at the level of events, which are structured representations of pieces of knowledge. Since there are normally multiple events in a sentence, the identification of discourse information at the event level can allow for a more detailed analysis of discourse elements than is possible when considering larger units of text, and can allow such constraints to be specified as search criteria in event extraction systems.

As discussed in chapters 3 and 4, our work on annotating discourse at the level of events involved defining a customised annotation scheme [8] that encodes various aspects of knowledge that can be relevant to discourse. This meta-knowledge annotation scheme has been used to enrich the GENIA event corpus of 1,000 biomedical abstracts with 36,858 events [29] to create the GENIA-MK corpus [9] and a corpus of 4 full papers pre-annotated with 1,710 GENIA events to create the FP-MK corpus [18]. The meta-knowledge annotation scheme can, in some respects, be compared roughly to the sentence-based classification schemes introduced above, in that it includes encoding of specific rhetorical functions. However, it differs in a number of ways. Firstly, the types of rhetorical functions encoded (referred to as *Knowledge Type)* in the meta-knowledge annotation scheme, e.g., fact, observation, analysis,

etc., are of a more abstract or high level nature than most of those used at the sentence level, which are often quite strongly tied to structural aspects of the article, with labels such as *background*, *experiment*, *conclusion*, etc. Secondly, further types of information that can be relevant to discourse analysis, e.g., certainty level, are also annotated for each event. As discussed in chapters 5-7, automatic recognition of different types of meta-knowledge for events has been demonstrated to be highly feasible [19-21].

Since the annotation of information about discourse function at the level of events has been shown to be complementary to sentence-based classification schemes [15], it is also likely that the same types of information could help to enrich previous efforts to annotate and recognise information about discourse structure and relations that use coarser-grained textual units (i.e., sentences and clauses). For example, considering patterns of discourse information at the event level could provide a more detailed account of the types of rhetorical moves that are made in text. In addition, considering the types of events that occur within the arguments of different types of discourse relations, or indeed annotating discourse relations between events, could complement previous efforts.

In this chapter, we describe our preliminary work on analysing the discourse structure of biomedical abstracts and full papers at the level of events. To our knowledge, this is a novel approach to event-level discourse analysis. Specifically, we look at patterns of transitions between events, in terms of knowledge type and certainty level, based on the event-level meta-knowledge annotations that are already present in the GENIA-MK and FP-MK corpora. Both types of information are relevant to understanding the structure or flow of discourse within documents. At the sen-

tence/clause level, it has been found previously that it is not possible to apply a fixed model of discourse structure consistently to all scientific texts [186, 202], and hence we also do not attempt to apply a fixed model at the level of events. Rather, we examine patterns of *Knowledge Type* and *Certainty Level* values that are assigned to sequences of events of various lengths. Firstly, we look at pairs of adjacent events, which facilitates an analysis of the local discourse contexts in which events appear. Secondly, for the GENIA-MK corpus, we examine the most common transition paths in abstracts, i.e., longer patterns of *Knowledge Type*/*Certainty Level* values that occur when extended chains of events are considered.

Due to the complexity of analysing the transitions between the values of all 5 meta-knowledge dimensions, and since not all of the dimensions are directly related to discourse structure (e.g., *Manner* encodes biologically-specific information), we consider only the two dimensions of the scheme that appear most relevant to the analysis of discourse structure, i.e. *Knowledge Type* and *Certainty Level.* Detailed descriptions of the meta-knowledge dimensions can be found in chapter 3. However, we provide below a brief summary/reminder of the *Knowledge Type* and *Certainty Level* dimensions:

Knowledge Type

This dimension captures the general information content of the event. Each event is classified into one of the following six categories:

- **Investigation**: Enquiries or investigations, which have either already been conducted or are planned for the future, typically marked by lexical cues like *examined, investigated* and *studied*, etc.

- **Observation**: Direct observations, often represented by lexical cues like *found* and *observed*, etc. Simple past tense sentences typically also describe observations.
- **Analysis**: Inferences, interpretations, speculations or other types of cognitive analysis, typically expressed by lexical cues like *suggest, indicate, therefore* and *conclude,* etc.
- **Fact:** General facts and well established knowledge, typically denoted by present tense event-triggers that describe biological processes, and are sometimes accompanied by the lexical cue *known*.
- **Method:** Events that describe experimental methods.
- **Other:** The default category, assigned to events that either do not fit into one of the above categories or do not express complete information.

Certainty Level

This dimension is only applicable to events whose *Knowledge Type* corresponds to Analysis. It encodes confidence in the truth of the event. Possible values are as follows:

- **L3**: No expression of uncertainty or speculation (default category).
- **L2**: High confidence or slight speculation. Typical markers include *suggest* and *indicate.*
- **L1**: Low confidence or considerable speculation; expressed using markers such as *may*, *might* and *perhaps*.

It is interesting to note that several of the lexical markers listed above have been used in psycholinguistic analysis of texts [203, 204]. However, in this case, lexical

items are grouped according to different psychological processes, e.g. SENSES (sense-related) and SOCIAL (social interaction). Guerini et al. [204] carried out a psycholinguistic analysis of scientific articles, and found that several lexical markers provided above can be used to determine the "virality" of scientific articles, i.e., how the language used within them affects how likely they are to be downloaded or bookmarked. The psycholinguistic classification of lexical items is different from the meta-knowledge classification shown above. For example, the psycholinguistic SENSES category contains verbs that denote different *Knowledge Type* categories, i.e., *Observation* and *Analysis*. Thus, the exact choice of certain lexical items in a paper may be motivated both in order to convey the right type of meta-knowledge and to ensure that the paper is as "viral" as possible.

The remainder of this chapter is structured as follows. In section 8.2, we look at the different types of transitions, both pairwise and paths, that occur in the abstracts of GENIA-MK corpus. In section 8.3, we examine the pairwise transitions in the full papers of the FP-MK corpus. Since there are sometimes significant differences in the distributions of meta-knowledge that occur in the different sections of full papers [18], the analysis of transitions in full papers is carried out in a section-wise manner.

8.2 Analysis of Meta-Knowledge Transitions in Abstracts

In this section we present a brief analysis of the meta-knowledge transitions observed in the GENIA-MK corpus of biomedical abstracts. We begin by examining patterns of individual, pair-wise transitions and then move on to look at transition paths (i.e., abstract level transition patterns).

8.2.1 Knowledge Type

8.2.1.1 Pair-wise Transitions

Figures 19-23 provide a summary of the pair-wise transitions **from** and **to** adjacent events in the GENIA-MK corpus, according to *Knowledge Type* categories. The blue lines represent the transitions **from** the category in focus (i.e., the category in the centre of the diagram), while the red lines indicate the transitions **to** that category. Similarly, the light blue boxes show the relative frequencies of each type of transition **from** the category, while the light red boxes show the relative frequencies of each type of transition **to** the category. The dotted lines represent reflexive transitions, i.e., cases where the *Knowledge Type* category of the adjacent event is the same as the event in focus. Transitions between all adjacent pairs of events are taken into account, i.e., not only those occurring within the boundaries of a sentence.

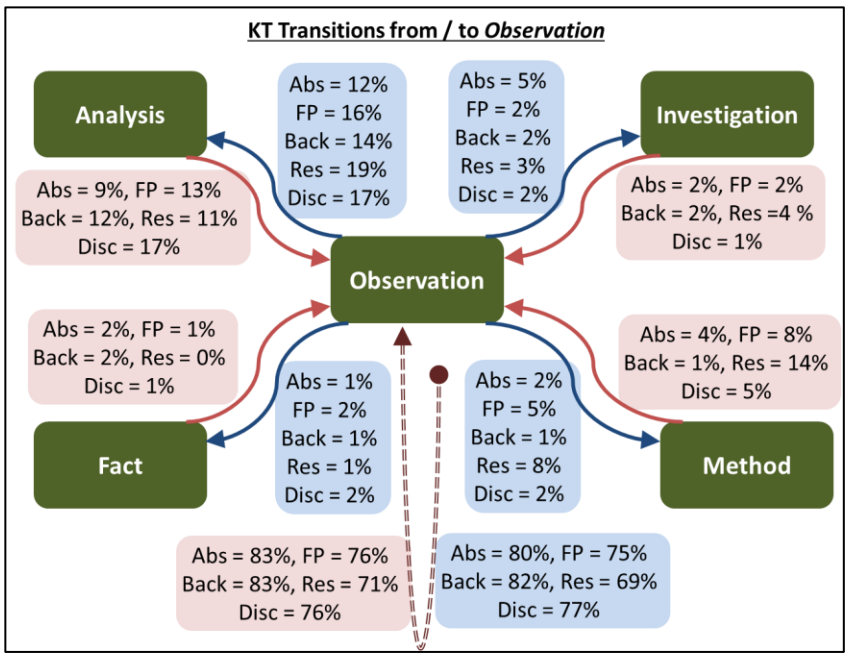

Figure 18. Transitions from / to *Knowledge Type* category *Observation* for Abstracts (Abs), Full Papers (FP), and the different sections within full papers, i.e., Background (Back), Results (Res), and Discussion (Disc)

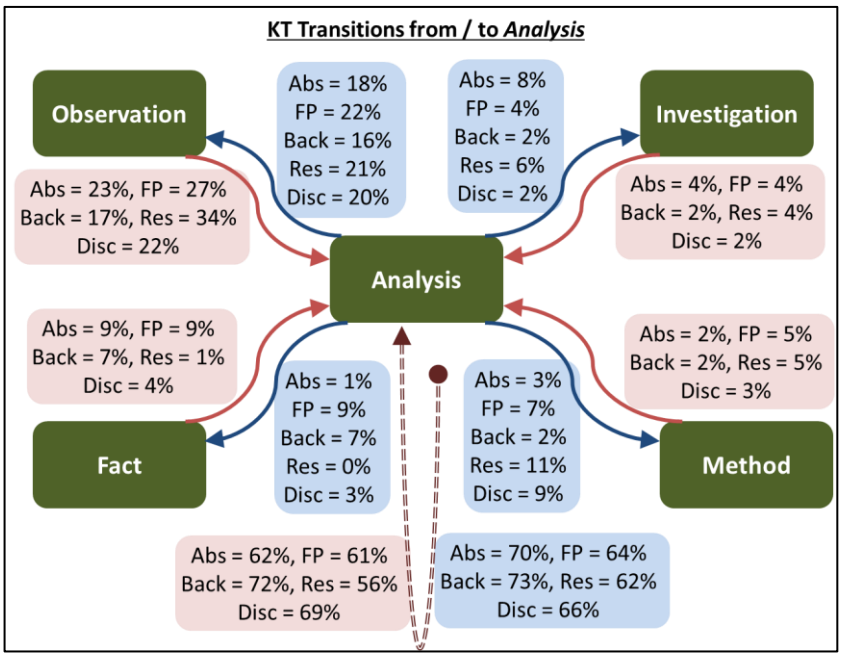

Figure 19. Transitions from / to *Knowledge Type* category *Analysis* for Abstracts (Abs), Full Papers (FP), and the different sections within full papers, i.e., Background (Back), Results (Res), and Discussion (Disc)

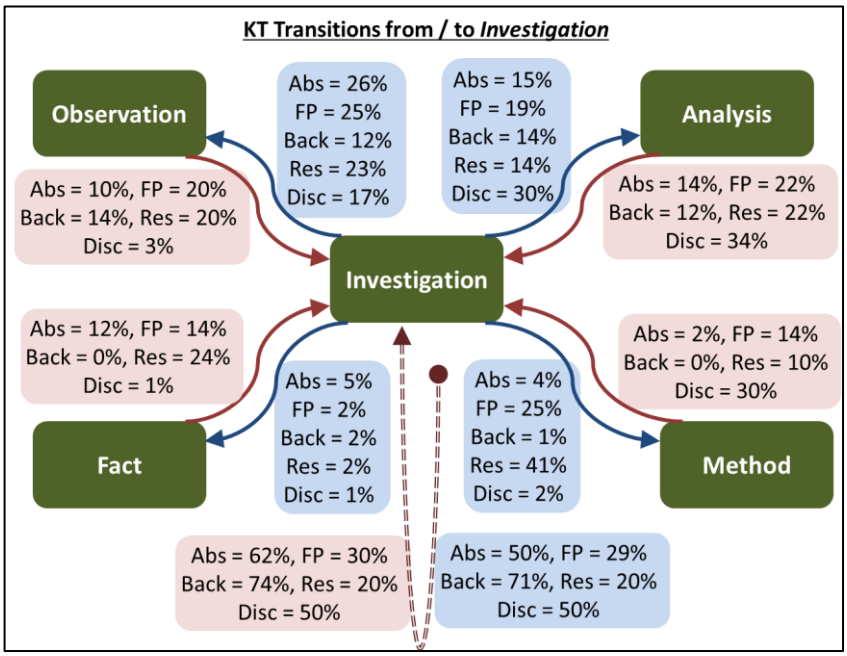

Figure 20. Transitions from / to *Knowledge Type* category *Investigation* for Abstracts (Abs), Full Papers (FP), and the different sections within full papers, i.e., Background (Back), Results (Res), and Discussion (Disc)

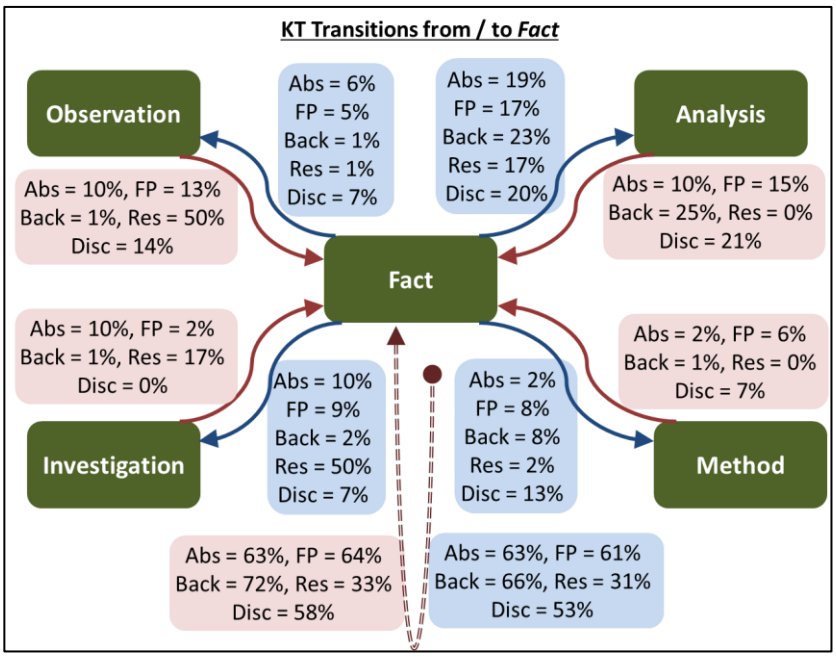

Figure 21. Transitions from / to *Knowledge Type* category *Fact* for Abstracts (Abs), Full Papers (FP), and the different sections within full papers, i.e., Background (Back), Results (Res), and Discussion (Disc)

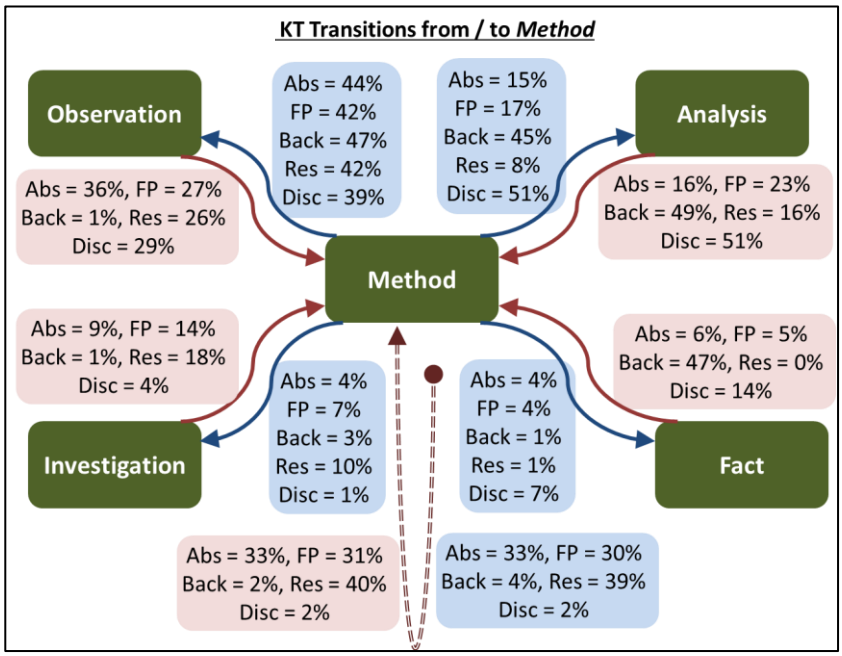

Figure 22. Transitions from / to *Knowledge Type* category *Method* for Abstracts (Abs), Full Papers (FP), and the different sections within full papers, i.e., Background (Back), Results (Res), and Discussion (Disc)

Observation

This is a highly reflexive category, with 80% of the transitions from *Observation* events leading to another *Observation* event; similarly 83% of all transitions to an *Observation* event originate from another *Observation* event. In terms of non-reflexive transitions, 12% of transitions originating from *Observation* events lead to *Analysis* events. This is because observations are often used as premises for analytical and hypothetical conclusions. Conversely, most of the non-reflexive transitions leading to *Observation* events start from *Analysis* events. This is probably because arguments presented in an abstract are often linked, i.e., the conclusion of an argument can be used as the premise of the next argument. A small but noticeable proportion (5%) of transitions starting from *Observation* events lead to *Investigation* events. However, in most cases, these observations are attributed to previous studies (as determined by the *Source* dimension of the meta-knowledge annotation scheme). That is, in these cases, a previous observation has been used as a premise for a new investigation.

Analysis

This is also a highly reflexive category, with 70% of the transitions from *Analysis* events leading to another *Analysis* event and 62% of transitions to *Analysis* events originating from other *Analysis* events. In terms of non-reflexive transitions, 18% of transitions from *Analysis* events lead to *Observation* events (possible reasons have been discussed above). Similarly, a significant proportion (23%) of transitions that lead to *Analysis* events start from *Observation* events. Transitions from *Analysis* events to events describing facts are very infrequent (1%). Conversely, 9% of all transitions leading to *Analysis* events originate from *Fact* events. This is because the

current state-of-the-art knowledge is sometimes analysed in order to situate or justify the study that is reported in a paper. Further evidence for this type of pattern is that a similar proportion (8%) of transitions starting from *Analysis* events lead to events describing investigations, i.e., in cases where background knowledge is stated and then analysed, it is usual that the next step is to use the results of the analysis as a basis for introducing the investigation to be carried out during the current study.

Investigation

This is a relatively less reflexive category, with only 50% of transitions from *Investigation* events leading to other *Investigation* events, and 62% of all transitions to *Investigation* events originating from other *Investigation* events. This is probably because the structure of abstracts is often such that only the main investigation is discussed at the beginning of the abstract, followed by observations and analyses. This argument is further supported by the fact that a significant number of transitions from *Investigation* events lead to either *Observation* (26%) or *Analysis* (15%) events.

Fact

This is also a less reflexive category: 63% of all transitions from *Fact* events lead to other *Fact* events, and vice versa. Events describing facts are often followed by events describing analyses (19%), mainly due to the reasons described in the *Analysis* section above. In some cases, facts serve as direct premises for investigations (10%). Less frequently, facts are directly followed by observations (6%).

Method

Only 33% of transitions from/to method events are reflexive. This is mainly because, in abstracts, authors tend to mention the methods used in their work only briefly (if at all). Since it is a natural progression for authors to move from the description of methods to the description of experimental results achieved according to the application of these methods, this explains why the highest proportion of transitions from *Method* events (44%) lead to *Observation* events. However, since the reporting of experimental outcomes or conclusions is of vital importance in abstracts, it is sometimes the case that observations themselves will be omitted, and authors will move straight from describing methods to analysing their findings. This goes towards explaining why 15% of *Method* events are directly followed by *Analysis* events. Most of the non-reflexive transitions that lead to *Method* events originate from *Observation* (36%). This can be explained by the fact that authors frequently present findings from previous studies in order to set the scene for introducing their own experimental methods. A significant percentage of transitions to *Method* events are from *Analysis* events (16%). There are a number of possible reasons for this. In some cases, an analysis of previous findings may be necessary in order to correctly justify the author's own methods. In other cases, authors may complete their discussion of one set of experiments and then move on to introducing a further set of methods.

Expected Values for Random Transitions

Considering the distribution of *Knowledge Type* categories in the GENIA-MK corpus (section 4.3), there is a 35% probability that the destination of a random transition will be an event with the *Knowledge Type* category of *Observation*. The next

most likely destination is an *Analysis* event. However, the probability of this transition is only 18%. Similarly, the probabilities of random transitions to the remaining *Knowledge Type* categories are even lower: 8% for *Fact*, 5% for *Investigation* and 3% for *Method*. These values indicate that the reflexivity of all *Knowledge Type* categories is much higher than what would be expected by chance. Therefore, there is a strong likelihood that two contiguous events will belong to the same *Knowledge Type* category. This information could be potentially useful for automatic identification of *Knowledge Type* categories.

The frequencies of observed transitions from *Observation* to *Fact* and from *Analysis* to *Fact* are significantly lower than what would be expected by chance. The frequencies of transitions from *Investigation* to *Fact* and from *Method* to *Fact* are also slightly less than the expected frequencies. This is mainly because the *Fact* events are almost eight times more reflexive than what would be expected by chance. Moreover, a majority of abstracts start with a *Fact* event. This is further explained in the following section.

8.2.1.2 Abstract Level Patterns

We examined the *Knowledge Type* values of the first and last event in each abstract in the GENIA-MK corpus. The results of this analysis are summarised in Table 31. In the majority of cases, authors begin by stating known facts as a scene-setting device for introducing their own work. The use of *Knowledge Type* categories other than *Fact* at the start of abstracts is considerably less frequent, with *Analysis* and *Observation* being the next most common categories. Analysis of the *Source* dimension of these event types reveals that they often pertain to previous studies, indicating that a discussion of previous findings is a common way to start the abstract.

Sometimes, scene-setting steps are missed out altogether, and the abstract launches directly into an explanation of the investigation to be undertaken. In very rare cases, even subject of investigation is missing, and the abstract gets straight down to the business of explaining the experimental setup and methodology.

Knowledge Type Category	**Abstracts Starting With**	**Abstracts Ending With**
Observation	10%	15%
Analysis	23%	78%
Investigation	9%	4%
Fact	54%	1%
Method	4%	2%

Table 31. Relative frequencies of abstracts starting and ending with events of each *Knowledge Type* category

Transition Pattern	**% in Abstracts**
Fact → Analysis → Observation → ... → Analysis	14%
Fact → Investigation → Observation → ... → Analysis	10%
Fact → Observation → ... → Analysis	8%
Analysis → Observation → ... → Analysis	7%
Analysis → Fact → Observation → ... → Analysis	6%
Analysis → Investigation → Observation → ... → Analysis	4%

Table 32. Key transition patterns for *Knowledge Type* values in abstracts and their frequencies

In the vast majority of cases, authors end their abstracts with an *Analysis* event, presenting a summary or interpretation of the most important findings of the experiments undertaken. However, there is a significant proportion of cases (15%) in which the abstract ends with an *Observation* event. This can happen when a significant experimental observation has occurred during the current study. Very occasionally, the abstracts end by presenting an investigative topic or method that the authors have identified for further exploration.

Although extended transition patterns of *Knowledge Type* values vary significantly in biomedical abstracts, we were able to identify several general patterns that occur with noticeable frequencies. Table 32 shows some of these transition patterns, along with the percentage of abstracts in which these patterns manifest themselves. Almost a quarter of all abstracts start with known facts, followed by analyses of previous work or a description of the investigation to be carried out in the current study; this is in turn followed by a description of experimental observations, and the abstract ends with an analysis of these observations. Interestingly, over 8% of the abstracts exhibit a simplified variant of this pattern, where the second transition to *Analysis* or *Investigation* is omitted and a direct link is made between the previously known facts and the (new) observations made by the authors. A possible explanation of this could be the need for brevity resulting from the fact that abstract size constraints vary between biomedical journals. A significant number of abstracts follow a slightly different *Knowledge Type* transition pattern. They start with an analysis of previous studies, followed by observations from the current study, and end with an analysis of findings. Variants of this pattern, which include a transition to a *Fact*, to help

to contextualise the analyses of previous studies, or present an *Investigation* between the first *Analysis* and *Observation* events, are also found in 10% of abstracts.

The above patterns suggest that while most biomedical abstracts loosely follow the *Creating A Research Space (CARS)* model proposed by Swales [194], a significant proportion of abstracts skip the first step of "establishing a territory", and assume that the reader is already familiar with the context. This could be partly due to the specialised nature of many biomedical journals.

8.2.2 Certainty Level

8.2.2.1 Pair-wise Transitions

Figures 24-26 summarise the pair-wise transitions **from** and **to** adjacent events in the GENIA-MK corpus, according to the *Certainty Level* category assigned to them. Similarly to figures above, the blue lines represent the transitions **from** the category in focus (i.e., the category in the centre of the diagram), while the red lines indicate the transitions **to** that category.

L3

This is a highly reflexive category, partly due to its high frequency of occurrence – 92% of events in the GENIA-MK corpus are expressed with the highest/default certainty level. In terms of non-reflexive transitions, 6% of transitions from *L3* events lead to *L2* events, and only 1% to *L1* events. As explained earlier, most abstracts start with a brief mention of previous knowledge (observations, analyses or facts), followed by a summary of investigations and the resulting observations, and conclude with analyses of experimental findings, which are often hedged.

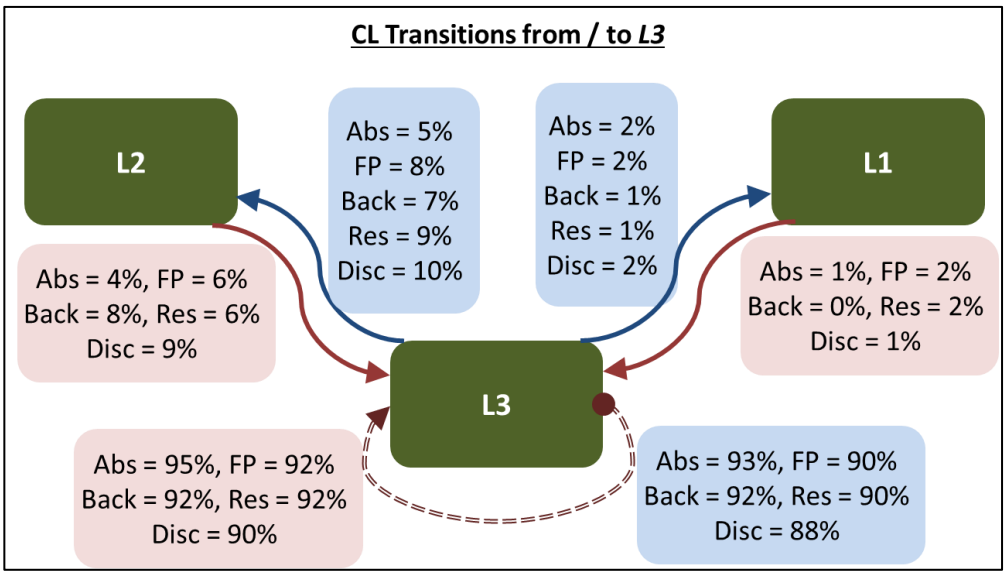

Figure 23. Transitions from / to *Certainty Level* category *L3* for Abstracts (Abs), Full Papers (FP), and the different sections within full papers, i.e., Background (Back), Results (Res), and Discussion (Disc)

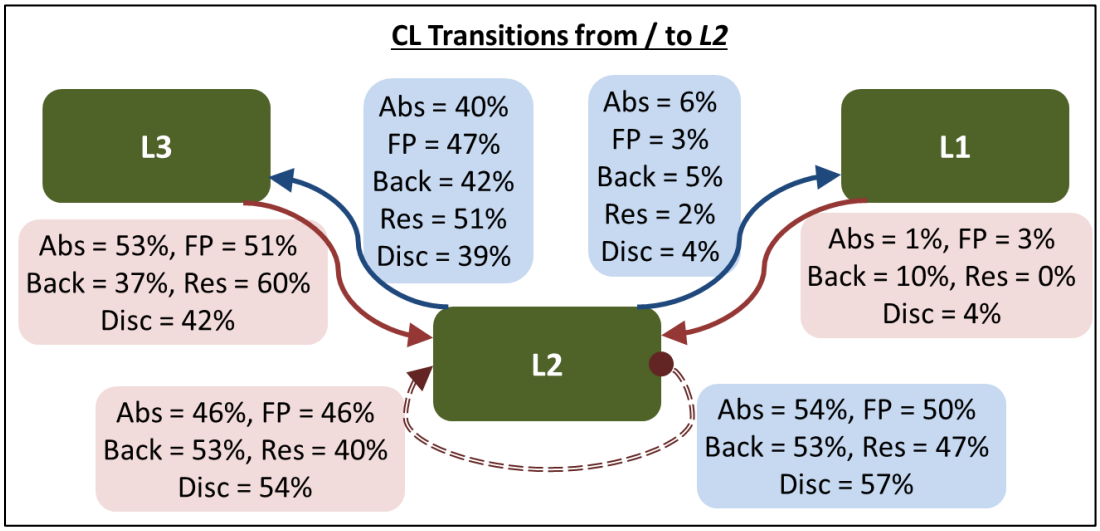

Figure 24. Transitions from / to *Certainty Level* category *L2* for Abstracts (Abs), Full Papers (FP), and the different sections within full papers, i.e., Background (Back), Results (Res), and Discussion (Disc)

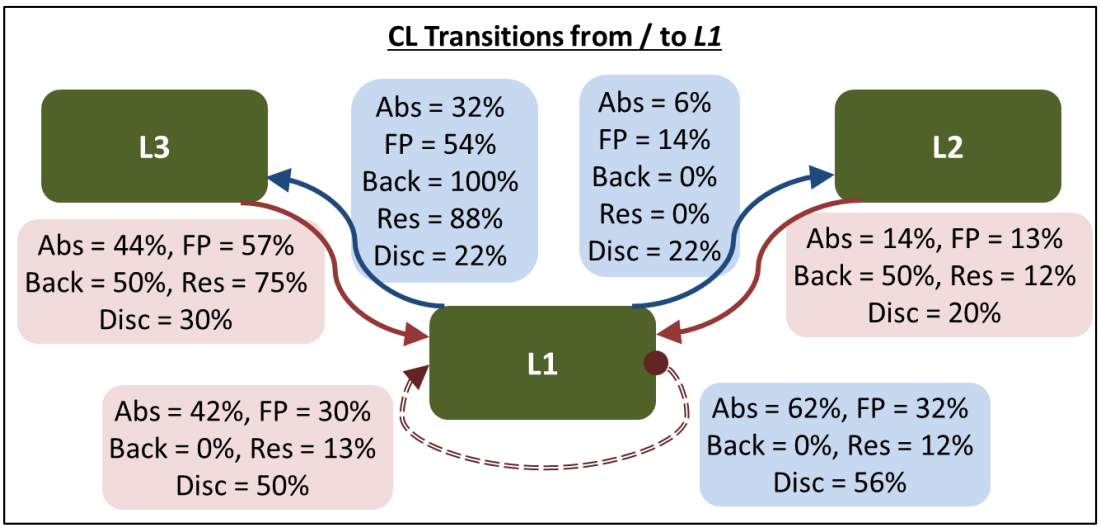

Figure 25. Transitions from / to *Certainty Level* category *L1* for Abstracts (Abs), Full Papers (FP), and the different sections within full papers, i.e., Background (Back), Results (Res), and Discussion (Disc)

L2

This is the least reflexive category, partly due to the fairly small number of *L2* events in the corpus as a whole. Also, since authors do not want to throw too much doubt on their findings, they are likely to avoid long chains of speculated events. This would also explain why a significant proportion (40%) of transitions from *L2* events leads back to *L3* events. This pattern occurs mostly in cases where, having described a set of observations (mostly *L3* events) and the corresponding analyses (probably *L2* events), the authors proceed to describe a different set of observations. Interestingly, 6% of transitions from *L2* events lead to *L1* events. These are mostly the cases where slightly hedged analyses are followed by bolder (highly speculative) extensions and corollaries.

L1

For similar reasons to *L2*, this is also a less reflexive category. Although a significant proportion of transitions from *L1* events lead to *L3* (34%) and *L2* (6%) events, the volumes of *L1* events are so small (less than 1% of all events) that they only account for around 1% of all transitions to *L3* and *L2*.

Expected Values for Random Transitions

The probability of a random transition to an *L3* event is 92%. The similar probabilities for *L2* and *L1* events are 6% and 2%, respectively. This indicates that the reflexivity values of *L2* and *L1* are many magnitudes higher than the corresponding expected values. Furthermore, the frequency of observed transitions from *L2* to *L1* is three times higher than the expected value. Similarly, the frequency of observed transitions from *L1* to *L2* is twice the expected value. This shows that the likelihood of an event being speculative significantly increases if the previous event is also speculative. This information is potentially very useful for automated identification of speculated events.

8.2.2.2 Abstract Level Patterns

We examined the *Certainty Level* values of the first and last event in each abstract in the GENIA-MK corpus. The results of this analysis are summarised in Table 33. As mentioned earlier, almost all abstracts start with either known facts, or previous observations, analyses, or investigations, i.e., events expressed with absolute certainty of occurrence (*L3*). However, although most abstracts end with analyses, authors will usually aim to have as much impact as possible at the end of abstract, so that readers are encouraged to look further into the main body of the text. Thus, if the

authors are sufficiently confident about the conclusions they have drawn about their experimental results, they will be stated without any degree of hedging. Due to the desire of authors not to hedge more than necessary (especially in abstracts), most hedged events that occur at the end of abstracts are only slightly hedged. A smaller, but still important percentage of terminal events are marked as highly speculative, as sometimes authors want to make a large impact by presenting possible analyses which, while highly speculative, are also highly innovative or controversial.

Certainty Level Category	**Abstracts Starting With**	**Abstracts Ending With**
L1	0%	19%
L2	1%	36%
L3	99%	45%

Table 33. Relative frequencies of abstracts starting and ending with events of each *Certainty Level* category

We observed that 28% of abstracts contain no speculated events, which reinforces the claim that authors will only introduce uncertainty into abstracts where absolutely necessary. Of the remaining abstracts, a significant majority (58%) include the following transition pattern: *L3* → *L2*. These are the cases where authors deploy slight hedging on the analyses of their findings. Sometimes, this pattern is repeated 2 or 3 times, mostly when abstracts report on multiple sets of observations, each followed by its corresponding analysis. We also observed that a small proportion of abstracts (5%) contain the following transition pattern: *L3* → *L2* → *L1*. As mentioned earlier, these are the cases where slightly hedged analyses are followed by bolder analyses, predictions or hypotheses, which can be a useful tool in helping to pique the reader's

curiosity. Interestingly, we also discovered that a significant proportion of abstracts (14%) contain the following transition pattern: *L3* → *L1*, i.e., observations and confident analyses are followed directly by highly speculated analyses or hypotheses.

8.3 Analysis of Meta-Knowledge Transitions in Full Papers

In this section we present a brief analysis of the meta-knowledge transitions observed in the FP-MK corpus. We have analysed the transitions from one *Knowledge Type*/*Certainty Level* category to another in each of three main sections: *Background*, *Results*, and *Discussion*.

8.3.1 Knowledge Type

Figures 19-23 (above) show the summary of pair-wise transitions **from** and **to** adjacent events in the FP-MK corpus, according to *Knowledge Type* categories. They include separate statistics for each of the main sections in full papers (i.e., *Background*, *Results*, and *Discussion*), as well as for the papers as a whole.

Observation

Overall distributions of transitions from and to *Observation* events in full papers are similar to those in abstracts. However, the reflexivity of *Observation* events is slightly less in full papers than in abstracts. This is partly because of the greater numbers of links between observations and analyses in full papers. The proportion of transitions from *Observation* events to *Analysis* events is significantly higher in full papers than in abstracts. This is because full papers contain many more observations, most of which are subsequently further analysed. This kind of linking between observations and analyses is particularly frequent in the *Results* and *Discussion* sections of full papers. Full papers contain slightly fewer transitions from *Observation*

to *Investigation*. This is mainly because the relative frequency of *Investigation* events is considerably lower in full papers than in abstracts.

Analysis

Full papers contain significantly more transitions from *Analysis* to *Fact* events, especially in *Background* and *Discussion* sections. This is because the stringent size constraints imposed for abstracts are relaxed for the body of full papers, and thus authors have greater opportunity to relate their work to the state-of-the-art in their domain. The overall reflexivity of *Analysis* events is slightly less in full papers than in abstracts. This is despite the fact that the overall relative frequency of *Analysis* events in full papers is higher than in abstracts. This can be explained by the more complex interweaving of analytical statements with observations or facts that is often found in full papers. The transitions from *Analysis* to *Observation* are much higher in full papers than in abstracts. This can again be explained by the more complex patterns of discourse shifts that occur in full papers, where multiple observations are introduced and analysed, in order to guide and convince the reader of the final conclusions that are drawn. Such patterns have particularly high frequency in the *Results* and *Discussion* sections of papers. Finally, full papers contain significantly fewer transitions from *Analysis* to *Investigation*. This is mainly because *Investigation* events rarely occur in some sections of full papers, whereas many abstracts contain a small number of *Investigation* events.

Investigation

Overall reflexivity of *Investigation* events in full papers is significantly less than in abstracts. As mentioned earlier, this is due to a lower relative frequency of *Investigation* events in full papers. Full papers contain significantly higher numbers of transi-

tions from *Investigation* events to *Method* events. Interestingly, almost all of these transitions are in the *Results* sections. This is probably due to the need to explain how particular aspects of the investigation were carried out by applying particular experimental methods. A similar percentage of transitions can be observed between *Method* and *Observation* events in the *Results* sections, showing that the next step is often to describe how the use of the method led to particular experimental observations. Full papers contain fewer transitions from *Investigation* events to *Fact* events than abstracts, once again due to the lower relative frequency of *Investigation* events in full papers. Full papers also contain slightly more transitions from *Investigation* events to *Analysis* events, especially in *Discussion* sections, where a direct link is made between the investigations undertaken and the findings resulting from them.

Fact

Overall distributions are similar to abstracts, with one minor difference: full papers contain more transitions from *Fact* to *Method*, especially in *Background* and *Discussion* sections. This is mainly because sometimes, authors make a direct link between background facts and the experimental methods used, omitting the intermediary link to investigations. This is especially the case when authors have already mentioned the investigations earlier in the text.

Method

We found no significant differences in the distribution of *Method* events in full papers and abstracts. This is partly due to the scarcity of *Method* events (in both GENIA-MK and FP-MK corpora) caused by the definition of bio-event used to annotate these corpora. As mentioned earlier, according to this definition, bio-events

are “dynamic bio-relations”. Most mentions of experimental methods do not constitute dynamic relations, and hence are not annotated as events in these corpora.

8.3.2 Certainty Level

Figures 24-26 (above) show the summary of pair-wise transitions **from** and **to** adjacent events in the FP-MK corpus, according to *Certainty Level* categories. They include separate statistics for each of the main sections in full papers (i.e., *Background*, *Results*, and *Discussion*), as well as for the papers as a whole.

L3

The distributions of transitions from/to *L3* events in full papers are similar to those in abstracts, except for one main difference: Full papers contain slightly more transitions from *L3* to *L2* events. This is probably due to the more detailed analytical discussion often found in full papers. Moreover, in the longer text of the body of the paper, authors tend to express more speculation than in abstracts. This is because, unlike in abstracts, where the main aim of authors is to try to sell the results of their research, the body of the paper provides much greater opportunity for analysis and discussion. Indeed, it would seem highly unusual if authors did not specify any uncertainty whilst analysing their results. The percentage of *L3* to *L2* transitions is highest in the *Results* sections of the full papers. Authors may thus be confident about some of their results, but not so confident about others. The percentage of these transitions drops in the *Discussion* section, suggesting that authors take a more confident tone in analysing their most definite results, in order to convince the reader of the reliability of their conclusions.

L2

Full papers contain slightly more transitions from *L2* to *L3* events. This is mainly due to the more frequent occurrence of contiguous observation-analysis transitions. Full papers contain significantly fewer transitions from *L2* to *L1* events. As mentioned above, such transitions are often made in abstracts for increased effect or impact, in order to grab the attention of the reader. It thus seems reasonable that fewer such transitions would occur in the body of the paper. If too many bold or controversial statements are made, readers may question the integrity of the study.

L1

Overall reflexivity of *L1* events is much lower in full papers than in abstracts. Although the relative frequency of L1 events is higher in full papers than in abstracts, they are more thinly spread out in full papers. The greater the number of highly speculative events that occur in sequence, the more wary the reader is likely to become. Thus, L1 events occur sporadically and are usually interspersed with more confident events to lessen their potentially negative impact. According to the previous observation, full papers contain a greater number of transitions from *L1* events to *L2* and *L3* events than abstracts.

8.4 Conclusion

We have investigated discourse patterns that occur in biomedical abstracts and full papers. In contrast to previous work on discourse structure, our analysis was conducted at the level of bio-events. Additionally, we have considered not only discourse/rhetorical functions (i.e., *Knowledge Type*) but also the certainty level. We used the GENIA-MK corpus of abstracts and the FP-MK corpus of full papers, both

containing meta-knowledge enriched event annotations, as the source of our analyses. We examined a number of different types of discourse patterns. For both abstracts and full papers, we considered patterns of pairwise transitions between events, considering *Knowledge Type* and *Certainty Level* separately. We explained probable reasons for our findings, and compared the results obtained for abstracts and full papers, revealing that there are a number of subtle and significant differences in the patterns of local discourse-level shifts that are observed within them. For abstracts, we additionally considered the complete transition paths (from the beginning of an abstract to its end) for *Knowledge Type* and *Certainty Level* values. This analysis showed that whilst there are some clear patterns of *Knowledge Type* and *Certainty Level* transitions in abstracts, these are by no means standard. Furthermore, we discovered that while most biomedical abstracts follow a generic model of rhetoric/information moves, authors often skip certain moves, assuming that the reader is already familiar with the context.

Chapter 9: Conclusion

In this chapter, we evaluate the progress against research objectives and hypotheses established at the beginning of the project. We also discuss the main areas of future work.

9.1 Evaluation of Research Objectives and Hypotheses

As discussed in section 1.2, four specific research objectives were established at the beginning of the project. The end-of-project evaluation of these objectives and hypotheses is as follows:

9.1.1 Objective # 1

O1 *To develop an annotation scheme for capturing the information necessary for the correct interpretation of bio-events*

Evaluation

We achieved this objective by developing the event-level meta-knowledge annotation scheme for capturing the necessary information required for the correct interpretation of bio-events. We also developed detailed annotation guidelines.

Peer Review and Verification

The initial proposal for the annotation scheme was presented at the Seventh International Conference on Language Resources and Evaluation (LREC 2010) [8]. We made minor modifications to the annotation scheme based on further analysis and feedback from reviewers and peers. The updated annotation scheme was evaluated through a case study. The results were presented at the ACL Workshop on Negation

and Speculation in Natural Language Processing (NeSp-NLP 2010) [13]. The final version of the annotation scheme was presented at the CLARIN/DARIAH Workshop on Automated Motif Discovery in Cultural Heritage and Scientific Communication Texts (AMICUS 2010) [14, 17].

9.1.2 Objective # 2

O_2 *To develop manually annotated corpora of bio-events with the required interpretative information*

Evaluation

We achieved this objective by developing two manually annotated corpora of bio-events enriched with meta-knowledge information, i.e., the GENIA-MK and FP-MK corpora. We trained two independent annotators from different backgrounds (one biology expert and one linguistics expert) to perform meta-knowledge annotations. The GENIA-MK corpus was created by adding meta-knowledge annotations to bio-events in the GENIA Event corpus, which comprises 1,000 biomedical abstracts containing 36,858 bio-events. The FP-MK corpus was created by adding meta-knowledge annotations to the bio-events in 4 full papers from the BioNLP'11 ST corpus, which contains 1,710 bio-events.

Peer Review and Verification

The results of the annotation project to enrich the GENIA Event corpus with meta-knowledge information (i.e., the creation of GENIA-MK corpus) were published in the journal BMC Bioinformatics [9]. The results of the annotation project to create the FP-MK corpus and the comparison of meta-knowledge annotations in abstracts

and full papers were presented at the Third LREC Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM 2012) [18].

9.1.3 Objective # 3

O_3 *To develop automated systems for enriching bio-events with the required interpretative information*

Evaluation

We achieved this objective by developing automated systems for identification of three meta-knowledge dimensions, i.e., polarity, manner and knowledge source. All three systems achieved high precision, recall and F-scores.

Peer Review and Verification

The results of our research work on the analysis and identification of bio-event polarity were published in the journal BMC Bioinformatics [19]. The automated system for the identification of event manner was presented at the Eighth International Conference on Language Resources and Evaluation (LREC 2012) [20]. The results of our work on the analysis and identification on knowledge source in bio-events were presented at the 14th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2013) [21].

9.1.4 Evaluation of Research Hypotheses

As stated in section 1.2, this research project had two main hypotheses:

H_1 *Discrete information about event interpretation can be identified – Meta-knowledge annotation can be performed at the event level*

H_2	*The above information can be automatically extracted – Meta-knowledge can be extracted automatically*

Evaluation

Through successful completion of the research objectives (above), we have proved that both of the above hypotheses are true.

9.2 Future Work

The main avenues of future work are as follows:

9.2.1 Meta-knowledge Annotation for Other Domains

The identification and extraction of events can be important in many different domains of academic and business analysis. However, the exact nature and definition of the events to be recognised will be specific to the domain. For most types of texts, the recognition of meta-knowledge will be a relevant sub-task of the event extraction process, since the textual context of events will always affect their interpretation, no matter what domain is being considered. Although we designed our meta-knowledge annotation scheme with a particular focus on the biomedical domain, the scheme is general enough to be suitable for application to other domains with some modifications.

Fellow researchers at the National Centre for Text Mining (NaCTeM) are currently investigating the feasibility of applying our meta-knowledge annotation scheme in the ISHER project[4], which aims to enhance search over digitised social history resources, through text mining-based rich semantic metadata extraction for collection

[4] http://www.nactem.ac.uk/DID-ISHER/

indexing, clustering and classification, thus supporting semantic search. This semantic metadata includes both named entities and events. The Automatic Content Evaluation (ACE) 2005 corpus [24] is being used as part of the training data. This corpus contains socio-political events, such as *Conflict* (indicated by verbs like *attack*, *demonstrate*, etc.) and *Justice* (indicated by verbs like *arrest*, *jail*, *sentence*, *fine*, etc.).

The research work will involve enriching relevant events in the corpus with meta-knowledge annotation. Preliminary results have shown that three of the original meta-knowledge dimensions are directly applicable to ACE i.e., *Polarity, Source* and *Certainty Level*, as these dimensions and their values represent general characteristics of all text types. *Manner* is not relevant to the social history domain but *Knowledge Type* is, although a different set of values may be required for each different domain, given that our current set of values are based on the types of information present in scientific research papers.

9.2.2 Meta-knowledge Extraction

As mentioned earlier, meta-knowledge extraction modules can be added to the state-of-the-art event extraction systems. This will allow the creation of more sophisticated systems that will be able to retrieve events with specified values of meta-knowledge. Such systems will allow researchers to carry out much more focussed searches over large bodies of text. The users of such systems will be able to retrieve documents containing events of a specified type, with specified participants, and also with specified values of meta-knowledge. For example, a user will be able to formulate a query to retrieve all documents containing *negated* (polarity) instances

of *intense* (manner) *Positive Regulation* (event type) of *p21ras proteins* (theme) by *IL-2* (cause) mentioned as a *tentative* (certainty level) *analysis* (knowledge type).

9.2.3 Discourse Analysis

The initial results (chapter 8) from investigations into event-based discourse analysis of scientific texts are encouraging. The scope of this investigation can be broadened by incorporating more varied types of bio-events and the remaining meta-knowledge dimensions (i.e., *Polarity*, *Knowledge Source* and *Manner*). The meta-knowledge transition patterns within each section of full papers should also be investigated. Furthermore, with the help of the BioDRB corpus, an investigation can be launched into whether there are correlations between particular types of discourse relations and the meta-knowledge values of the events that occur within the argument text spans of these relations. This could provide additional features to improve the accuracy of systems designed to recognise discourse relations automatically. Finally, a further line of enquiry is the investigation of event-level discourse analysis in other knowledge / research domains, such as social history.

References

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

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

3. Cohen, A.M., Hersh, W.R.: A survey of current work in biomedical text mining. Brief Bioinform **6** (2005) 57-71

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

5. Ananiadou, S., Kell, D.B., Tsujii, J.: Text mining and its potential applications in systems biology. Trends Biotechnol **24** (2006) 571-579

6. Ding, J., Berleant, D., Nettleton, D., Wurtele, E.: Mining MEDLINE: abstracts, sentences, or phrases. Proceedings of the Pacific Symposium on Biocomputing 7, Lihue, Hawaii, USA (2002) 326-337

7. Ananiadou, S., Nenadic, G.: Automatic Terminology Management in Biomedicine. In: Ananiadou, S., McNaught, J. (eds.): Text Mining for Biology and Biomedicine. Artech House, London / Boston (2006) 67-98

8. Nawaz, R., Thompson, P., McNaught, J., Ananiadou, S.: Meta-Knowledge Annotation of Bio-Events. 7th International Conference on Language Resources and Evaluation (LREC-2010), Malta (2010) 2498-2507

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

10. Oda, K., Kim, J.D., Ohta, T., Okanohara, D., Matsuzaki, T., Tateisi, Y., Tsujii, J.: New challenges for text mining: mapping between text and manually curated pathways. BMC Bioinformatics **9** (2008) S5

11. Hull, D., Pettifer, S., Kell, D.: Defrosting the Digital Library: Bibliographic Tools for the Next Generation Web. PLoS Computational Biology **4** (2008) e1000204

12. Dagan, I., Glickman, O., Magnini, B.: The PASCAL Recognising Textual Entailment Challenge. In: Quiñonero-Candela, J.D., I.; Magnini, B.; d'Alché-Buc, F. (ed.): Machine Learning Challenges, Vol. 3944. Springer (2006) 177-190

13. Nawaz, R., Thompson, P., Ananiadou, S.: Evaluating a Meta-Knowledge Annotation Scheme for Bio-Events. Workshop on Negation and Speculation in Natural Language Processing (NeSp-NLP 2010), ACL 2010, Uppsala, Sweden (2010) 69-77

14. Nawaz, R., Thompson, P., Ananiadou, S.: Event Interpretation: A Step towards Event-Centred Text Mining. First International Workshop on Automated Motif Discovery in Cultural Heritage and Scientific Communication Texts (AMICUS 2010), CLARIN/DARIAH 2010, Vienna, Austria (2010)

15. Liakata, M., Thompson, P., de Waard, A., Nawaz, R., Maat, H.P., Ananiadou, S.: A Three-Way Perspective on Scientific Discourse Annotation for Knowledge Extraction. Proceedings of the Workshop on Detecting Structure in Scholorly Discourse (DSSD) (2012) 37-46

16. Batista-Navarro, R., Kontonatsios, G., Mihăilă, C., Thompson, P., Nawaz, R., Korkontzelos, I., Ananiadou, S.: Supporting Discourse Phenomena in an Interoperable NLP Framework. CICLing 2013, Samos, Greece (2013)

17. Ananiadou, S., Thompson, P., Nawaz, R.: Improving Search Through Event-based Biomedical Text Mining. First International Workshop on Automated Motif Discovery in Cultural Heritage and Scientific Communication Texts (AMICUS 2010), CLARIN/DARIAH 2010, Vienna, Austria (2010)

18. Nawaz, R., Thompson, P., Ananiadou, S.: Meta-Knowledge Annotation at the Event Level: Comparison between Abstracts and Full Papers. Proceedings of the Third LREC Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM 2012) (2012) 24-21

19. Nawaz, R., Thompson, P., Ananiadou, S.: Negated Bio-events: Analysis and Identification. BMC Bioinformatics **14** (2013)

20. Nawaz, R., Thompson, P., Ananiadou, S.: Identification of Manner in Bio-Events. Eighth International Conference on Language Resources and Evaluation (LREC 2012) (2012) 3505-3510

21. Nawaz, R., Thompson, P., Ananiadou, S.: Something Old, Something New: Identifying Knowledge Source in Bio-Events. CICLing 2013, Samos, Greece (2013)

22. Nawaz, R., Thompson, P., Ananiadou, S.: Towards Event-Based Discourse Analysis of Biomedical Text. CICLing 2013, Samos, Greece (2013)

23. Sauri, R., Pustejovsky, J.: FactBank: A Corpus Annotated with Event Factuality. Language Resources and Evaluation **43** (2009) 227-268

24. NIST: Automatic Content Extraction 2005 Evaluation (ACE05). Vol. 2013 (2007)

25. Akhmatova, E.: Textual Entailment Resolution via Atomic Propositions. First Challenge Workshop on Recognizing Textual Entailment, Southhampton, UK (2005)

26. Hickl, A., Bensley, J.: A Discourse Commitment-Based Framework for Recognizing Textual Entailment. ACL-COLING Workshop on Textual Entailment and Paraphrasing (WTEP 2007), Prague (2007)

27. Harris, Z.: The structure of science information. Journal of Biomedical Informatics **35** (2002) 215-221

28. Friedman, C., Kra, P., Rzhetsky, A.: Two biomedical sublanguages: a description based on the theories of Zellig Harris. Journal of Biomedical Informatics **35** (2002) 222-235

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

30. 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.: Gene Ontology: tool for the unification of biology. Nature Genetics **25** (2000) 25-29

31. Pyysalo, S., Ginter, F., Heimonen, J., Bj¨orne, J., Boberg, J., J¨arvinen, J., Salakoski, T.: BioInfer: A Corpus for Information Extraction in the Biomedical Domain. BMC Bioinformatics **8** (2007)

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

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

34. Kim, J.D., Pyysalo, S., Ohta, T., Bossy, R., Nguyen, N., Tsujii, J.: Overview of BioNLP Shared Task 2011. BioNLP Shared Task 2011 Workshop (2011) 1-6

35. Blaschke, C., Andrade, M.A., Ouzous, C., Valencia, A.: Automatic extraction of biological information from scientific text: Protein-protein interactions. Intelligent Systems for Molecular Biology (1999) 60-67

36. Leitner, F., Mardis, S.A., Krallinger, M., Cesareni, G., Hirschman, L.A., Valencia, A.: An Overview of BioCreative II.5. IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM **7** (2010) 385--399

37. Korbel, J., Doerks, T., Jensen, L.J., Perez-Iratxeta, C., Kaczanowski, S., Hooper, S., Andrade, M., Bork, P.: Systematic Association of Genes to Phenotypes by Genome and Literature Mining. PLoS Biology **3** (2005)

38. Sam, L., Mendonça, E., Li, J., Blake, J., Friedman, C., Lussier, Y.: PhenoGO: an integrated resource for the multiscale mining of clinical and biological data. BMC Bioinformatics **10 Suppl 2** (2009) S8

39. Chun, H.W., Tsuruoka, Y., Kim, J.D., Shiba, R., Nagata, N., Hishiki, T., Tsujii, J.: Extraction of gene-disease relations from Medline using domain dictionaries and machine learning. Pac Symp Biocomput (2006) 4-15

40. Ozgur, A., Vu, T., Erkan, G., Radev, D.R.: Identifying gene-disease associations using centrality on a literature mined gene-interaction network. Bioinformatics **24** (2008) i277-285

41. Chatr-aryamontri, A., Ceol, A., Palazzi, L.M., Nardelli, G., Schneider, M.V., Castagnoli, L., Cesareni, G.: MINT: the Molecular INTeraction database. Nucleic acids research **35** (2007) 572-574

42. Bader, G.D., Cary, M.P., Sander, C.: Pathguide: a Pathway Resource List. Nucleic Acids Research **34** (2006) 504-506

43. Camon, E., Magrane, M., Barrell, D., Lee, V., Dimmer, E., Maslen, J., Binns, D., Harte, N., Lopez, R., Apweiler, R.: The Gene Ontology Annotation (GOA) Database: Sharing Knowledge in Uniprot with Gene Ontology. Nucleic Acids Research **32** (2004.) 262–266

44. Ananiadou, S., Pyysalo, S., Tsujii, J., Kell, D.: Event extraction for systems biology by text mining the literature. Trends in Biotechnology **28** (2010) 381-390

45. Hirschman, L., Blaschke, C.: Evaluation of Text Mining in Biology. In: Ananiadou, S., McNaught, J. (eds.): Text Mining for Biology and Biomedicine. Artech House, Boston / London (2006) 213-245

46. Pyysalo, S., Ohta, T., Kim, J.-D., Tsujii, J.: Static Relations: a Piece in the Biomedical Information Extraction Puzzle. Workshop on Natural Language Processing in Biomedicine (BioNLP) - NAACL 2009, Boulder, Colorado (2009) 1-9

47. Kim, J.-D., Ohta, T., Pyysalo, S., Kano, Y., Tsujii, J.: Overview of BioNLP'09 Shared Task on Event Extraction. ACL Workshop on BioNLP: Shared Task, Boulder, Colorado, USA (2009) 1–9

48. Kim, J.-D., Wang, Y., Takagi, T., Yonezawai, A.: Overview of Genia Event Task in BioNLP Shared Task 2011. BioNLP 2011, Association for Computational Linguistics: Human Language Technologies, Portland, Oregon, USA (2011) 7-15

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

50. Beisswanger, E., Lee, V., Kim, J.J., Rebholz-Schuhmann, D., Splendiani, A., Dameron, O., Schulz, S., Hahn, U.: Gene Regulation Ontology (GRO): design principles and use cases. 21st International Congress of the European Federation for Medical Informatics (MIE 2008), Gothenburg, Sweden (2008) 9-14

51. Buyko, E., Beisswanger, E., Hahn, U.: The GeneReg Corpus for Gene Expression Regulation Events - An Overview of the Corpus and its In-Domain and Out-of-Domain Interoperability. 7th International Conference on Language Resources and Evaluation (LREC-2010), Malta (2010)

52. Hirschman, L., Yeh, A., Blaschke, C., Valencia, A.: Overview of BioCreAtIvE: critical assessment of information extraction for biology. BMC Bioinformatics **6** (2005) S1

53. Hersh, W., Cohen, A., Ruslen, L., Roberts, P.: TREC 2007 Genomics track overview. Proceedings of the Sixteenth Text REtrieval Conference (2007)

54. Kim, J.-D., Ohta, T., Tsuruoka, Y., Tateisi, Y., Collier, N.: Introduction to the bio-entity recognition task at JNLPBA. International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA) (2004) 70-75

55. Nedellec, C.: Learning language in logic - genic interaction extraction challenge. 4th Learning in Logic Workshop (LLL05) (2005) 31-37

56. Kim, J.-D., Nguyen, N., Wang, Y., Tsujii, J., Takagi, T., Yonezawa, A.: The Genia Event and Protein Coreference tasks of the BioNLP Shared Task 2011. BMC Bioinformatics **13** (2012) S1

57. Pyysalo, S., Ohta, T., Rak, R., Sullivan, D., Mao, C., Wang, C., Sobral, B., Tsujii, J.i., Ananiadou, S.: Overview of the ID, EPI and REL tasks of BioNLP Shared Task 2011. BMC Bioinformatics **13** (2012) S2

58. Bossy, R., Jourde, J., Manine, A.-P., Veber, P., Alphonse, E., van de Guchte, M., Bessieres, P., Nedellec, C.: BioNLP Shared Task - The Bacteria Track. BMC Bioinformatics **13** (2012) S3

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

60. Miwa, M., Saetre, R., Kim, J.D., Tsujii, J.: Event extraction with complex event classification using rich features. J Bioinform Comput Biol **8** (2010) 131-146

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

62. Bjorne, J., Heimonen, J., Ginter, F., Airola, A., Pahikkala, T., Salakoski, T.: Extracting Complex Biological Events with Rich Graph-Based Feature Sets. ACL Workshop on BioNLP: Shared Task, Boulder, Colorado, (2009) 10–18

63. Björne, J., Ginter, F., Salakoski, T.: University of Turku in the BioNLP'11 Shared Task. BMC Bioinformatics **13** (2012) S4

64. Björne, J., Ginter, F., Pyysalo, S., Tsujii, J., Salakoski, T.: Scaling up Biomedical Event Extraction to the Entire PubMed. ACL Workshop on Biomedical Natural Language Processing. Association for Computational Linguistics, Uppsala, Sweden (2010) 28-36

65. Riedel, S., McClosky, D., Surdeanu, M., McCallum, A., Manning, C.: Model combination for event extraction in BioNLP 2011. BioNLP Shared Task 2011 Workshop. Association for Computational Linguistics (2011) 51-55

66. Kemper, B., Matsuzaki, T., Matsuoka, Y., Tsuruoka, Y., Kitano, H., Ananiadou, S., Tsujii, J.i.: PathText: a text mining integrator for biological pathway visualizations. Bioinformatics **26** (2010) i374-i381

67. Miyao, Y., Ohta, T., Masuda, K., Tsuruoka, Y., Yoshida, K., Ninomiya, T., Tsujii, J.i.: Semantic Retrieval for the Accurate Identification of Relational Concepts in Massive Textbases. ACL '06: 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL. Association for Computational Linguistics (2006) 1017-1024

68. Hara, T., Miyao, Y., Tsujii, J.: Adapting a probabilistic disambiguation model of an HPSG parser to a new domain. Proceedings of IJCNLP, Vol. 3651 (2005) 199-210

69. Tsuruoka, Y., Tsujii, J.: Bidirectional Inference with the Easiest-First Strategy for Tagging Sequence Data. Proceedings of HLT/EMNLP 2005 (2005) 467-474

70. Miyao, Y., Sagae, K., Saetre, R., Matsuzaki, T., Tsujii, J.: Evaluating contributions of natural language parsers to protein-protein interaction extraction. Bioinformatics **25** (2009) 394-400

71. Tsai, R.T., Chou, W.C., Su, Y.S., Lin, Y.C., Sung, C.L., Dai, H.J., Yeh, I.T., Ku, W., Sung, T.Y., Hsu, W.L.: BIOSMILE: a semantic role labeling system for biomedical verbs using a maximum-entropy model with automatically generated template features. BMC Bioinformatics **8** (2007) 325

72. Venturi, G., Montemagni, S., Marchi, S., Sasaki, Y., Thompson, P., Mcnaught, J., Ananiadou, S.: Bootstrapping a Verb Lexicon for Biomedical Information Extraction. Proceedings of CICLING. Springer-Verlag, Mexico City, Mexico (2009) 137-148

73. Cohen, K.B., Palmer, M., Hunter, L.: Nominalization and alternations in biomedical language. PLoS ONE **3(9)** (2008)

74. Sagae, K., Tsujii, J.i.: Dependency parsing and domain adaptation with LR models and parser ensembles. Proceedings of the CoNLL 2007 Shared Task (2007)

75. Heiner, M., Koch, I., Will, J.: Model validation of biological pathways using Petri nets - demonstrated for apoptosis. Biosystems **75** (2004) 15-28

76. Kell, D., Oliver, S.: Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era. Bioessays **26** (2004) 99-105

77. Luciano, J.S., Stevens, R.D.: e-Science and biological pathway semantics. BMC Bioinformatics **8** (2007) S3

78. Ye, Y., Doak, T.G.: A parsimony approach to biological pathway reconstruction/inference for genomes and metagenomes. PLoS Computational Biology **5** (2009)

79. Ananiadou, S., Kell, D.B., Tsujii, J.: Text mining and its potential applications in systems biology. Trends in Biotechnology **24** (2006) 571-579

80. Kell, D.B.: Metabolomics, modelling and machine learning in systems biology: towards an understanding of the languages of cells. The 2005 Theodor Bücher Lecture. Febs J **273** (2006) 873-894

81. Spasic, I., Simeonidis, E., Messiha, H.L., Paton, N.W., Kell, D.B.: KiPar, a tool for systematic information retrieval regarding parameters for kinetic modelling of yeast metabolic pathways. Bioinformatics **25** (2009) 1404-1411

82. Herrgard, M.J., Swainston, N., Dobson, P., Dunn, W.B., Arga, K.Y., Arvas, M., Bluthgen, N., Borger, S., Costenoble, R., Heinemann, M., Hucka, M., Le Novere, N., Li, P., Liebermeister, W., Mo, M.L., Oliveira, A.P., Petranovic, D., Pettifer, S., Simeonidis, E., Smallbone, K., Spasic, I., Weichart, D., Brent, R., Broomhead, D.S., Westerhoff, H.V., Kirdar, B., Penttila, M., Klipp, E., Palsson, B.O., Sauer, U., Oliver, S.G., Mendes, P., Nielsen, J., Kell, D.B.: A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. Nat Biotechnol **26** (2008) 1155-1160

83. Rzhetsky, A., Iossifov, I., Koike, T., Krauthammer, M., Kra, P., Morris, M., Yu, H., Duboué, P.A., Weng, W., Wilbur, W.J.: GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathway data. Journal of Biomedical Informatics **37** (2004) 43-53

84. Rajagopalan, D., Agarwal, P.: Inferring pathways from gene lists using a literature-derived network of biological relationships. Bioinformatics **21** (2005) 788-793

85. Santos, C., Eggle, D., States, D.: Wnt pathway curation using automated natural language processing: combining statistical methods with partial and full parse for knowledge extraction. Bioinformatics **21** (2005) 1653-1658

86. Tsuruoka, Y., Tsujii, J., Ananiadou, S.: FACTA: a text search engine for finding associated biomedical concepts. Bioinformatics **24** (2008) 2559-2560

87. Nobata, C., Cotter, P., Okazaki, N., Rea, B., Sasaki, Y., Tsuruoka, Y., Tsujii, J., Ananiadou, S.: Kleio: a knowledge-enriched information retrieval system for biology. 31st Annual International ACM SIGIR, Singapore (2008) 787-788

88. Tsuruoka, Y., Miwa, M., Hamamoto, K., Tsujii, J., Ananiadou, S.: Discovering and visualizing indirect associations between biomedical concepts. Bioinformatics **27** (2011) i111–i119

89. Blaschke, C., Leon, E., Krallinger, M., Valencia, A.: Evaluation of BioCreAtIvE assessment of task 2. BMC Bioinformatics **6** (2005) S16

90. Lisacek, F., Chichester, C., Kaplan, A., Sandor, A.: Discovering paradigm shift patterns in biomedical abstracts: Application to neurodegenerative diseases. Proceedings of SMBM (2005) 212-217

91. de Waard, A., Shum, B., Carusi, A., Park, J., Samwald, M., Sándor, Á.: Hypotheses, Evidence and Relationships: The HypER Approach for Representing Scientific Knowledge Claims. Proceedings of the Workshop on Semantic Web Applications in Scientific Discourse (2009)

92. Yeh, A.S., Hirschman, L., Morgan, A.A.: Evaluation of text data mining for database curation: lessons learned from the KDD Challenge Cup. Bioinformatics **19** (2003) i331-i339

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

94. Medlock, B., Briscoe, T.: Weakly supervised learning for hedge classification in scientific literature. Proceedings of ACL (2007) 992-999

95. Hyland, K.: Talking to the academy: Forms of hedging in science research articles. Written Communication **13** (1996) 251-281

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

97. Rizomilioti, V.: Exploring Epistemic Modality in Academic Discourse Using Corpora. In: Arnó Macià, E., Soler Cervera, A., Rueda Ramos, C. (eds.): Information Technology in Languages for Specific Purposes. Springer, New York (2006) 53-71

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

99. Thompson, P., Venturi, G., McNaught, J., Montemagni, S., Ananiadou, S.: Categorising modality in biomedical texts. Proceedings of the LREC 2008 Workshop on Building and Evaluating Resources for Biomedical Text Mining, Marrakech, Morocco (2008) 27-34

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

101. Hyland, K.: Metadiscourse: Exploring interaction in writing. Continuum Intl Pub Group (2005)

102. Mizuta, Y., Korhonen, A., Mullen, T., Collier, N.: Zone analysis in biology articles as a basis for information extraction. International Journal of Medical Informatics **75** (2006) 468-487

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

104. Ruch, P., Boyer, C., Chichester, C., Tbahriti, I., Geissbühler, A., Fabry, P., Gobeill, J., Pillet, V., Rebholz-Schuhmann, D., Lovis, C.: Using argumentation to extract key sentences from biomedical abstracts. International Journal of Medical Informatics **76** (2007) 195-200

105. McKnight, L., Srinivasan, P.: Categorization of sentence types in medical abstracts. AMIA Annu Symp Proc. (2003) 440-444

106. Langer, H., Lungen, H., Bayerl, P.S.: Text type structure and logical document structure. Proceedings of the ACL Workshop on Discourse Annotation (2004)

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

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

109. de Waard, A., Maat, H.P.: A Classification of Research Verbs to Facilitate Discourse Segment Identification in Biological Text. Proceedings of the Interdisciplinary Workshop on Verbs: The Identification and Representation of Verb Features. (2010)

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

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

112. Pyysalo, S., Ginter, F., Heimonen, J., Björne, J., Boberg, J., Järvinen, J., Salakoski, T.: BioInfer: A Corpus for Information Extraction in the Biomedical Domain. BMC Bioinformatics **8** (2007)

113. Sanchez-Graillet, O., Poesio, M.: Negation of protein-protein interactions: analysis and extraction. Bioinformatics **23** (2007) i424-432

114. Rubin, V.L.: Stating with certainty or stating with doubt: Intercoder reliability results for manual annotation of epistemically modalized statements. Proceedings of NAACL-HLT (2007) 141-144

115. Hoye, L.: Adverbs and modality in English. Longman (1997)

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

117. Nawaz, R., Thompson, P., Ananiadou, S.: Evaluating a meta-knowledge annotation scheme for bio-events. Proceedings of the Workshop on Negation and Speculation in Natural Language Processing (2010) 69-77

118. Kim, J.D., Ohta, T., Oda, K., Tsujii, J.: From text to pathway: corpus annotation for knowledge acquisition from biomedical literature. In: Brazma, A., S., M., Akutsu, T. (eds.): Proceedings of the 6th Asia-Pacfific Bioinformatics Conference, Vol. 6. Imperial College Press (2008) 165-176

119. Genia, T.: XConc Suite. (2006)

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

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

122. Knight, J.: Null and Void. Nature **422** (2003) 554-555

123. Greenberg, J.H. (ed.): Universals of Human Language, Vol. 4: Syntax. Stanford University Press, Stanford, California (1978)

124. Tottie, G.: Negation in English Speech and Writing: A Study in Variation. Academic Press, New York (1991)

125. Horn, L.R.: A Natural History of Negation. CSLI, Stanford, CA (2001)

126. Mutalik, P.G., Deshpande, A., Nadkarni, P.M.: Use of General-purpose Negation Detection to Augment Concept Indexing of Medical Documents: A Quantitative Study Using the UMLS. Journal of the American Medical Informatics Association **8** (2001) 598 - 609

127. Olsen, B.: Journal of Negative Results in Biomedicine. Vol. 2011 (2002)

128. Smialowski, P., Pagel, P., Wong, P., Brauner, B., Dunger, I., Fobo, G., Frishman, G., Montrone, C., Rattei, T., Frishman, D., Ruepp, A.: The Negatome Database: A Reference Set of Non-Interacting Protein Pairs. Nucleic acids research **38** (2010) 540-544

129. Ceusters, W., Elkin, P., Smith, B.: Negative Findings in Electronic Health Records and Biomedical Ontologies: A Realist Approach. International Journal of Medical Informatics **76** (2007) 326-333

130. Garten, Y., Coulet, A., Altman, R.: Recent Progress in Automatically Extracting Information from the Pharmacogenomic Literature. Pharmacogenomics **11** (2010) 1467-1489

131. Krallinger, M.: Importance of Negations and Experimental Qualifiers in Biomedical Literature. Workshop on Negation and Speculation in Natural Language Processing (NeSp-NLP 2010), ACL 2010, Uppsala, Sweden (2010) 46-49

132. Kilicoglu, H., Bergler, S.: Syntactic Dependency Based Heuristics for Biological Event Extraction. BioNLP 2009 Workshop (2009) 119-127

133. Kim, J.-D., Ohta, T., Pyysalo, S., Kano, Y., Tsujii, J.: Overview of BioNLP'09 Shared Task on Event Extraction. Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task (2009) 1-9

134. MacKinlay, A., Martinez, D., Baldwin, T.: Biomedical event annotation with CRFs and precision grammars. BioNLP 2009 Shared Task. Association for Computational Linguistics, Boulder, Colorado (2009) 77-85

135. Van Landeghem, S., Saeys, Y., De Baets, B., Van de Peer, Y.: Analyzing text in search of bio-molecular events: a high-precision machine learning framework. BioNLP 2009 Shared Task. Association for Computational Linguistics, Boulder, Colorado (2009) 128-136

136. Sarafraz, F., Nenadic, G.: Using SVMs with the Command Relation Features to Identify Negated Events in Biomedical Literature. Workshop on Negation and Speculation in Natural Language Processing (NeSp-NLP 2010), ACL 2010, Uppsala, Sweden (2010) 78-85

137. Morante, R., Schrauwen, S., Daelemans, W.: Corpus-Based Approaches to Processing the Scope of Negation Cues: An Evaluation of the State of the Art. In: Bos, J., Pulman, S. (eds.): Ninth International Conference on Computational Semantics (IWCS 2011), Oxford, UK. (2011) 350--354

138. Vincze, V., Szarvas, G., Mora, G., Ohta, T., Farkas, R.: Linguistic Scope-based and Biological Event-based Speculation and Negation Annotations in the Genia Event and BioScope Corpora. Fourth International Symposium on Semantic Mining in Biomedicine (SMBM), Cambridgeshire, UK (2010)

139. Morante, R., Daelemans, W.: A Metalearning Approach to Processing the Scope of Negation. Thirteenth Conference on Computational Natural Language Learning (CoNLL). Association for Computational Linguistics, Boulder, Colorado (2009) 21-29

140. Harabagiu, S., Hickl, A., Lacatusu, F.: Negation, Contrast and Contradiction in Text Processing. Twenty-First National Conference on Artificial Intelligence (AAAI-06), Boston, MA (2006)

141. Huang, Y., Lowe, H.J.: A Novel Hybrid Approach to Automated Negation Detection in Clinical Radiology Reports. Journal of the American Medical Informatics Association **14** (2007) 304 - 311

142. Chapman, W.W., Bridewell, W., Hanbury, P., Cooper, G.F., Buchanan, B.B.: A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries. Journal of Biomedical Informatics **34** (2001) 301 - 310

143. Tolentino, H., Matters, M., Walop, W., Law, B., Tong, W., Liu, F., Fontelo, P., Kohl, K., Payne, D.: Concept Negation in Free Text Components of Vaccine Safety Reports. AMIA Annual Symposium (2006)

144. Elkin, P.L., Brown, S.H., Bauer, B.A., Husser, C.S., Carruth, W., Bergstrom, L.R., Wahner-Roedler, D.L.: A controlled trial of automated classification of negation from clinical notes. BMC Medical Informatics and Decision Making **5** (2005) 13

145. Morante, R.: Descriptive Analysis of Negation Cues in Biomedical Texts. Seventh International Language Resources and Evaluation (LREC 2010), Valletta, Malta (2010) 1429-1436

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

147. Agarwal, S., Yu, H.: Biomedical negation scope detection with Conditional Random Fields. Journal of the American Medical Informatics Association (JAMIA) **17** (2010) 696-701

148. Boytcheva, S., Strupchanska, A., Paskaleva, E., Tcharaktchiev, D., Str, D.G.: Some Aspects of Negation Processing in Electronic Health Records. International Workshop Language and Speech Infrastructure for Information Access in the Balkan Countries, Borovets, Bulgaria. (2005) 1-8

149. Averbuch, M., Karson, T.H., Ben-Ami, B., Maimond, O., Rokachd, L.: Context-Sensitive Medical Information Retrieval. 11th World Congress on Medical Informatics (MEDINFO-2004). IOS Press, San Francisco, CA (2004) 1-8

150. Goldin, I.M., Chapman, W.W.: Learning to Detect Negation with 'Not' in Medical Texts. ACM-SIGIR 2003 (2003)

151. Goryachev, S., Sordo, M., Zeng, Q.T., Ngo, L.: Implementation and Evaluation of Four Different Methods of Negation Detection. DSG (2006)

152. Rokach, L., Romano, R., Maimon, O.: Negation Recognition in Medical Narrative Reports. Information Retrieval **11** (2008) 499-538

153. Councill, I.G., McDonald, R., Velikovich, L.: What's Great and What's Not: Learning to Classify the Scope of Negation for Improved Sentiment Analysis. Workshop on Negation and Speculation in Natural Language Processing (NeSp-NLP 2010), ACL 2010, Uppsala, Sweden (2010) 51-59

154. Kim, J.-D., Ohta, T., Pyysalo, S., Kano, Y., Tsujii, J.: Overview of BioNLP'09 Shared Task on Event Extraction. ACL Workshop on BioNLP: Shared Task, Boulder, Colorado, USA (2009) 1–9

155. Morante, R.: Descriptive Analysis of Negation Cues in Biomedical Texts. Seventh International Language Resources and Evaluation Conference (LREC 2010), Valletta, Malta (2010) 1429-1436

156. Wiegand, M., Balahur, A., Roth, B., Klakow, D., Montoyo, A.: A Survey on the Role of Negation in Sentiment Analysis. Workshop on Negation and Spec-

ulation in Natural Language Processing (NeSp-NLP 2010), ACL 2010, Uppsala, Sweden (2010) 60-68

157. Wilson, T., Wiebe, J., Hoffmann, P.: Recognizing Contextual Polarity in Phrase-level Sentiment Analysis. Human Language Technologies Conference/Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005), Vancouver, Canada (2005) 347–354

158. Langacker, R.: On Pronominalization and the Chain of Command. In: Reibel, D., Schane, S. (eds.): Modern Studies in English. Prentice-Hall, Englewood Cliffs, NJ. (1969) 160-186

159. Reinhart, T.M.: The Syntactic Domain of Anaphora. Foreign Literatures and Linguistics, Vol. PhD. Massachusetts Institute of Technology (1976)

160. Barker, C., Pullum, G.K.: A Theory of Command Relations. Linguistics and Philosophy **13** (1990) 1-34

161. Dumais, S., Platt, J., Heckerman, D., Sahami, M.: Inductive Learning Algorithms and Representations for Text Categorization. In: Gardarin, G. (ed.): Information and knowledge management. ACM Press, Bethesda, MD (1998) 148-155

162. Escudero, G., Mhrquez, L., Rigau, G.: A Comparison between Supervised Learning Algorithms for Word Sense Disambiguation. 4th Conference on Computational Natural Language Learning, CoNLL'2000, Stroudsburg, PA, USA (2000) 31-36

163. Mitchell, T.: Machine Learning. McGraw Hill (1997)

164. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo, CA. (1993)

165. Quinlan, J.R.: Induction of Decision Trees. Machine learning **1** (1986) 81-106

166. Kingsford, C., Salzberg, S.L.: What are Decision Trees? Nature Biotechnology **26** (2008) 1011-1013

167. Breiman, L.: Random Forests. Machine Learning **45** (2001) 5-32

168. Chen, X.-W., Liu, M.: Prediction of Protein–Protein Interactions Using Random Decision Forest Framework. Bioinformatics **21** (2005) 4394-4400

169. Qi, Y., Klein-Seetharaman, J., Bar-Joseph, Z.: Random Forest Similarity for Protein-Protein Interaction Prediction from Multiple Sources. Pacific Symposium on Biocomputing (2005) 531-542

170. Zhang, H.: The Optimality of Naive Bayes. 17th International FLAIRS Conference Miami Beach, Florida (2004)

171. Cortes, C., Vapnik, V.: Support-Vector Networks. Machine learning **20** (1995) 273-297

172. Aha, D.W., Kibler, D., Albert, M.K.: Instance-Based Learning Algorithms. Machine learning **6** (1991) 37-66

173. Miyao, Y., Tsujii, J.: Feature Forest Models for Probabilistic HPSG Parsing. Computational Linguistics **34** (2008) 35-80

174. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. SIGKDD Explorations **11** (2009) 10-18

175. le Cessie, S., van Houwelingen, J.C.: Ridge Estimators in Logistic Regression. Applied Statistics. **41** (1992) 91-201

176. Platt, J.: Machines using Sequential Minimal Optimization. In: Schoelkopf, B., Burges, C., Smola, A. (eds.): Advances in Kernel Methods - Support Vector Learning. MIT Press (1998)

177. Wilson, T., Wiebe, J., Hwa., R.: Just How Mad Are You? Finding Strong and Weak Opinion Clauses. 21st Conference of the American Association for Artificial Intelligence (AAAI-2004) (2004) 761-769

178. Joshi, M., Penstein-Rose, C.: Generalizing Dependency Features for Opinion Mining. ACL-IJCNLP 2009 Conference, Suntec, Singapore (2009) 313-316

179. Caruana, R., Niculescu-Mizil, A.: An Empirical Comparison of Supervised Learning Algorithms. 23rd International Conference on Machine Learning, Pittsburgh, PA (2006) 161-168

180. Tsai, R., Chou, W.-C., Su, Y.-S., Lin, Y.-C., Sung, C.-L., Dai, H.-J., Yeh, I., Ku, W., Sung, T.-Y., Hsu, W.-L.: BIOSMILE: A semantic role labeling system for biomedical verbs using a maximum-entropy model with automatically generated template features. BMC Bioinformatics **8** (2007) 325

181. Palmer, M., Gildea, D., Kingsbury, P.: The Proposition Bank: An Annotated Corpus of Semantic Roles. Computational Linguistics **31** (2005) 71-106

182. Marcus, M.P., Santorini, B., Marcinkiewicz, M.A.: Building a large annotated corpus of English: The Penn Treebank. Computational Linguistics **19** (1994) 313-330

183. Thompson, P., McNaught, J., Montemagni, S., Calzolari, N., Del Gratta, R., Lee, V., Marchi, S., Monachini, M., Pezik, P., Quochi, V., Rupp, C.J., Sasaki, Y., Venturi, G., Rebholz-Schuhmann, D., Ananiadou, S.: The BioLexicon: a large-scale terminological resource for biomedical text mining. BMC Bioinformatics **12** (2011) 397

184. Morante, R., Sporleder, C. (eds.): Proceedings of the Workshop on Negation and Speculation in Natural Language Processing, Uppsala, Sweden (2010)

185. Liakata, M., Teufel, S., Siddharthan, A., Batchelor, C.: Corpora for Conceptualisation and Zoning of Scientific Papers. 7th International Conference on Language Resources and Evaluation (LREC-2010), Malta (2010)

186. Teufel, S.: Argumentative Zoning. University of Edinburgh, Edinburgh (1999)

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

188. Ananiadou, S., Pyysalo, S., Tsujii, J., Kell, D.B.: Event extraction for systems biology by text mining the literature. Trends Biotechnol **28** (2010) 381-390

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

190. Mullen, T., Mizuta, Y., Collier, N.: A baseline feature set for learning rhetorical zones using full articles in the biomedical domain. ACM SIGKDD Explorations **7** (2005) 52-58

191. Sandor, Å., de Waard, A.: Identifying Claimed Knowledge Updates in Biomedical Research Articles. Proceedings of the Workshop on Detecting Structure in Scholarly Discourse (DSSD) (2012) 10-17

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

193. Marcu, D., Echihabi, A.: An unsupervised approach to recognizing discourse relations. Proceedings of ACL. Association for Computational Linguistics (2002) 368-375

194. Swales, J.: Genre Analysis: English in Academic and Research Settings. Cambridge University Press (1990)

195. Carlson, L., Marcu, D., Okurowski, M.E.: Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory: Current and New Directions in Discourse and Dialogue. In: Kuppevelt, J., Smith, R.W. (eds.), Vol. 22. Springer Netherlands (2003) 85-112

196. Mann, W.C., Thompson, S.A.: Rhetorical structure theory: Toward a functional theory of text organization. Text **8** (1988) 243-281

197. Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., Webber, B.: The Penn Discourse TreeBank 2.0. Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC) (2008) 2961 - 2968

198. Prasad, R., McRoy, S., Frid, N., Joshi, A., Yu, H.: The biomedical discourse relation bank. BMC Bioinformatics **12** (2011) 188

199. Hajič, J., Smrž, O., Zemánek, P., Šnaidauf, J., Beška, E.: Prague Arabic Dependency Treebank: Development in Data and Tools. NEMLAR International Conference on Arabic Language Resources and Tools (2004) 110-117

200. Xue, N., Xia, F., Chiou, F., Palmer, M.: The Penn Chinese TreeBank: Phrase Structure Annotation of a Large Corpus. Natural Language Engineering **11** (2005) 207-238

201. Hajicova, E.: The Prague Dependency Treebank: Crossing the Sentence Boundary. The Second Workshop on Text, Speech, Dialogue (1999) 20-27

202. Duszak, A.: Academic discourse and intellectual styles. Journal of Pragmatics **21** (1994) 291-313

203. Pennebaker, J., Francis, M.: Linguistic inquiry and word count: LIWC. Erlbaum Publishers (2001)

204. Guerini, M., Pepe, A., Lepri, B.: Do linguistic style and readability of scientific abstracts affect their virality? : Sixth International AAAI Conference on Weblogs and Social Media (ICWSM) (2012) 475 - 478