Manchester Metropolitan University

Towards Event-based Discourse Analysis of Biomedical Text

RAHEEL NAWAZ, PAUL THOMPSON, AND SOPHIA ANANIADOU

University of Manchester, UK

ABSTRACT

Annotating biomedical text with discourse-level information is a well-studied topic. Several research efforts have annotated textual zones (e.g., sentences or clauses) with information about rhetorical status, whilst other efforts have linked and classified sets of text spans according to the type of discourse relation holding between them. A relatively new approach has involved annotating meta-knowledge (i.e., rhetorical intent and other types of information concerning interpretation) at the level of bio-events, which are structured representations of pieces of biomedical knowledge. In this paper, 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.

KEYWORDS: meta-knowledge, event, bio-event, discourse analysis.

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, conclusions, etc. The automatic identification of such information can help in tasks such as isolating new knowledge claims [1]. Within the biomedical domain, this information can in turn be useful for tasks such as maintaining models of biomedical processes [2] or the curation of biomedical databases [3].

Several annotation schemes, e.g., [4-6] have been developed to classify textual zones according to their rhetorical status or general information content. Such zones are usually not understood in isolation, but rather in relation to others [7]. Therefore, for certain tasks, such as automatic summarisation, it is important to gain a fuller understanding of how information conveyed in the text is arranged to form a coherent discourse. Work in this area has involved defining a model that describes the structure of the introductions to scientific articles [8] and examining patterns of argumentative zones that occur in scientific abstracts [9].

A further approach to discourse analysis has been to identify and characterise links between sentences and clauses. Several efforts to produce annotated corpora or automated systems have been based around the Penn TreeBank corpus of open domain news articles [10]. This corpus was enriched by [11] with discourse trees, based on Rhetorical Structure Theory (RST) [12], A system was created by [7] to predict certain classes of discourse relations automatically. The Penn Discourse TreeBank (PDTB) [13] added discourse relations to the Penn TreeBank, both implicit and explicit, that hold between pairs of text spans. The Biomedical Discourse Relation Bank (BioDRB) [14] annotates the same types of relations in biomedical research articles.

All of the studies above considered sentences or clauses as the units of annotation. In contrast, the present work is concerned with discourse information at the level of *events*, which are structured representations of pieces of knowledge. In particular, we focus on bio-events, which encode biological reactions or processes. The automatic identification of events can facilitate sophisticated semantic searching, allowing researchers to perform structured searches over events extracted from a large body of text [15].

The utility of events has resulted in the appearance of a number of event-annotated corpora in recent years, e.g., [16-18]. The shared tasks on event extraction at BioNLP workshops, e.g., [19] have helped to stimulate further research into event extraction. Since there are normally multiple events in a sentence, the identification of discourse infor-

mation at the event level can allow for a more detailed analysis of discourse elements than is possible when considering larger units of text.

Previous work on annotating discourse at the level of events has involved defining a customised annotation scheme [20] encoding various aspects of knowledge that can be relevant to discourse. This *meta-knowledge* scheme has been used to enrich the GENIA event corpus of 1,000 biomedical abstracts (36,858 events) [16] to create the GENIA-MK corpus [21], and a corpus of 4 full papers pre-annotated with 1,710 GENIA events to create the FP-MK corpus [22].

The meta-knowledge annotation scheme is somewhat comparable to the sentence-based classification schemes introduced above, in that it includes encoding of specific rhetorical functions, e.g., fact, observation, analysis (referred to as *Knowledge Type* (KT))*. However, further types of relevant to discourse analysis. e.g., certainty level (*CL)*, are also annotated for each event. Automatic recognition of different types of meta-knowledge for events has been demonstrated to be highly feasible [23, 24].

The annotation of information about discourse function at the level of events has been shown to be complementary to sentence-based classification schemes [25], meaning that event-based discourse analysis could help to enrich previous efforts to annotate and recognise discourse information using coarser-grained textual units.

In this paper, 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 *KT* and *CL*, based on the event-level meta-knowledge annotations that are already present in the GENIA-MK and FP-MK corpora. At the sentence/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 [9], and hence we also do not attempt this at the event level. Rather, we examine patterns of *KT* and *CL* values assigned to sequences of events of various lengths.

The remainder of this paper is structured as follows. In section 2, we provide further details about events and the meta-knowledge annotation scheme. In section 3, we look at the different types of transitions, both between pairs of adjacent events and for longer paths of events that occur in the abstracts of GENIA-MK corpus. In section 4, we examine the pairwise transitions in the full papers of the FP-MK corpus, while section 5 provides some concluding remarks and directions for future work.

TRIGGER:	*augmented*
TYPE:	**positive_regulation**
THEME:	***c-jun mRNA*** : RNA_molecule
CAUSE:	***LTB4*** : organic_molecule

Fig. 1. Typical representation of the bio-event contained in sentence S1

2 Bio-events and their Enrichment with Meta-knowledge

In this section, we provide a brief introduction to bio-events, and describe the meta-knowledge annotation scheme that has been designed to enrich them with additional information about their interpretation, including discourse-level information.

2.1 *Bio-events*

In its most general form, a **textual event** can be described as an action, relation, process or state expressed in the text [26]. More specifically, it is a structured semantic representation of a piece of information contained in the text. Events are usually anchored to text fragments that are central to the description of the event, e.g., *event-trigger*, *event-participants* and *event-location*, etc. A number of corpora of general language with event-like annotations have been produced, e.g., [27, 28].

A **bio-event** is a specialised textual event, constituting a dynamic bio-relation involving one or more participants [16]. These participants can be bio-entities or (other) bio-events, and are each assigned a semantic role like *theme* and *cause*. Bio-events and bio-entities are also typically assigned semantic types/classes from particular taxonomies/ontologies. Consider the sentence S1: "*We conclude that LTB4 may augment c-jun mRNA*". This sentence contains a single bio-event of type *positive_regulation*, which is anchored to the verb *augmented*. Figure 1 shows a typical structured representation of this bio-event, with two participants: *c-jun mRNA* and *LTB4*, which have been assigned semantic types and roles within the event.

2.2 *Meta-Knowledge*

Whilst Figure 1 shows the typical information that would be extracted from sentence S1 by an event extraction system, there is other infor-
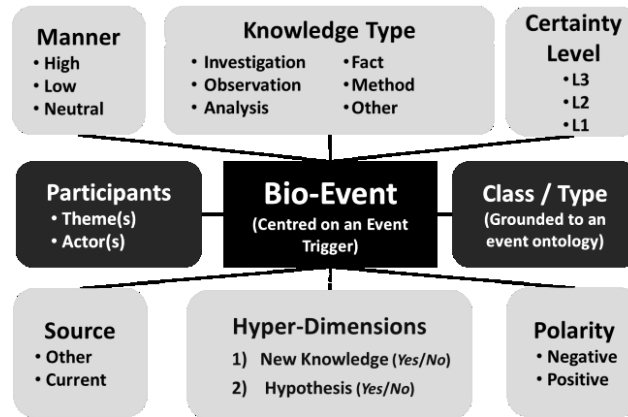
Fig. 2. Meta-knowledge annotation scheme

mation present in S1 that must be extracted if the event is to be interpreted correctly. For example, in terms of *KT*, the event does not represent a definite fact, but rather an analytical conclusion drawn by the authors. Similarly, the presence of the word *may* shows that the conclusion drawn is a tentative one, i.e., the *CL* of the analysis encoded by the event is low. The meta-knowledge annotation scheme (Figure 2) is able to capture this information about the event. The scheme consists of 5 different meta-knowledge dimensions, which encode not only discourse-relevant information, but also other common types of information that are necessary for the correct interpretation of a bio-event.

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, we consider only the two dimensions of the scheme that are most relevant in this respect, i.e. *KT* and *CL*. These are defined as follows:

Knowledge Type (KT)

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.
− **Observation**: Direct experimental observations
− **Analysis**: Inferences, interpretations, speculations or other types of analysis.
− **Fact:** General facts and well established knowledge.

- **Method:** Events that describe experimental methods.
- **Other:** Default category, assigned to events that either do not fit into one of the above categories or do not express complete information.

Certainty Level (CL)

This dimension is only applicable to events whose KT 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.
- **L1**: Low confidence or considerable speculation.

3   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. We begin with patterns of individual, pair-wise transitions and then move on to look at longer transition paths.

3.1   *Knowledge Type (KT)*

Pair-wise Transitions

Figure 3 provides a summary of the pair-wise transitions **from** and **to** adjacent events in the GENIA-MK corpus, according to *KT* categories. The black lines represent the transitions **from** the category in the centre of the diagram), while the grey lines indicate the transitions **to** that category. Similarly, the dark grey boxes show the relative frequencies of each type of transition **from** the category, while the light grey boxes show the relative frequencies of each type of transition **to** the category. The dotted lines boxes surrounded by dotted lines represent reflexive transitions, i.e., cases where the *KT* 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.

Observation: This is a highly reflexive category, with 80% of transitions from *Observation* leading to another *Observation*; similarly 83% of transitions to an *Observation* originate from another *Observation*. In
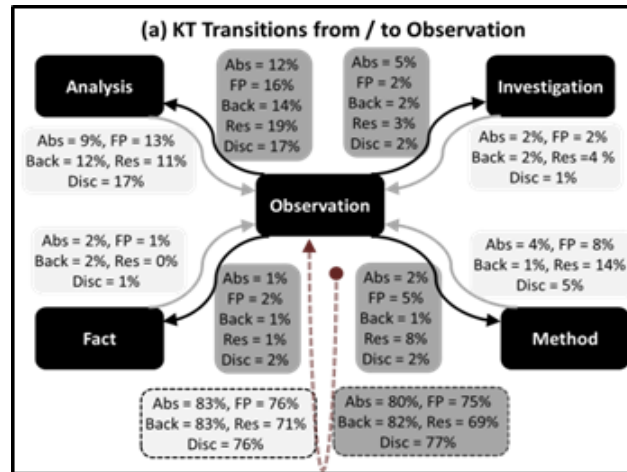
Fig. 3. Transitions from/to *KT* categories for Abstracts (Abs), Full Papers (FP), and the different sections within full papers, i.e., Background (Back), Results (Res), and Discussion (Disc). Continued to the next page.

terms of non-reflexive transitions, 12% of transitions originating from *Observation* lead to *Analysis*, because observations are often used as premises for analytical and hypothetical conclusions. Conversely, most non-reflexive transitions leading to *Observation* start from *Analysis*. This is probably due to the linked nature of arguments presented in an abstract, 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* lead to *Investigation*. However, in most cases, these observations are attributed to previous studies (as determined by the *Source* dimension of the annotation scheme). That is, 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* leading to another *Analysis* and 62% of transitions to *Analysis* originating from *Analysis*. In terms of non-reflexive transitions, 18% of transitions from *Analysis* lead to *Observation* (possible reasons have been discussed above). Similarly, a significant proportion (23%) of transitions that lead to *Analysis* start from *Observation*. Transitions from *Analysis to Fact* are very infrequent (1%). Conversely, 9% of all transitions leading to *Analysis* originate from *Fact*. This is because the state-of-the-art knowledge is sometimes analysed in
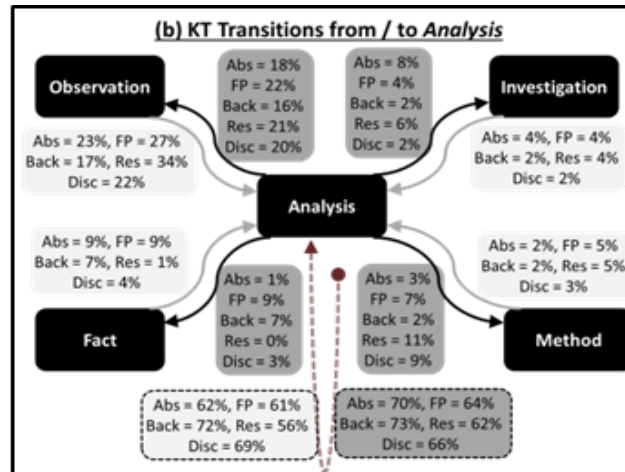
Fig. 3, continued. Continued to the next page.

order to situate or justify the study that is reported in a paper. Further evidence for this pattern is that a similar proportion (8%) of transitions starting from *Analysis* lead to *Investigation*. i.e., in cases where background knowledge is stated and analysed, it is usual that the analysed information is used as a basis for introducing the focussed investigation of the current study.

Investigation: This is a less reflexive category, with only 50% of transitions from *Investigation* leading to other *Investigation*s, and 62% transitions to *Investigation* events originating from other *Investigation*s. This is because the main investigation is usually discussed only at the beginning of the abstract, followed by observations and analyses. This argument is further supported the significant number of transitions from *Investigation* that lead to *Observation* (26%) or *Analysis* (15%).

Fact: This is also a less reflexive category: 63% of all transitions from *Fact* lead to other *Fact*s, and vice versa. *Fact*s are often followed by *Analysis* (19%), as described in the *Analysis* section above. In some cases, *Fact*s serve as direct premises for *Investigation* (10%). Infrequently, *Fact*s are directly followed by *Observation*s (6%).

Method: Only 33% of transitions from/to *Method* are reflexive. In abstracts, authors tend to mention the methods used in their work only briefly (if at all). Since it is natural for authors to move from the de-
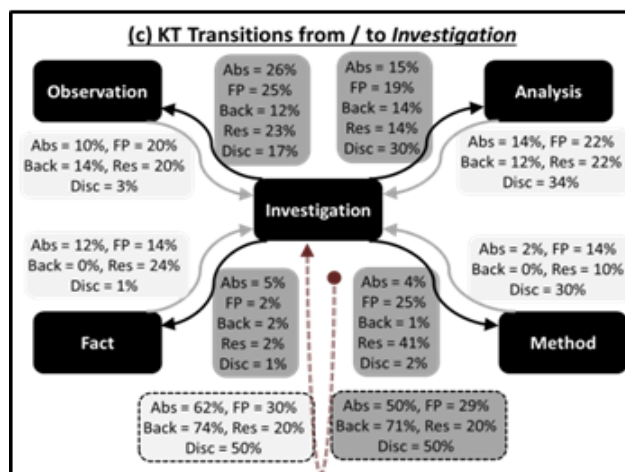
Fig. 3, continued. Continued to the next page.

scription of methods to subsequent experimental results, 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, observations will sometimes be omitted, and authors move straight from describing methods to analysing their findings. This goes towards explaining why 15% of *Method*s are directly followed by *Analysis*. Most of the non-reflexive transitions that lead to *Method* originate from *Observation* (36%). This is because authors frequently present findings from previous studies to set the scene for introducing their own experimental methods. A significant percentage of transitions to *Method* are from *Analysis* (16%). In some cases, an analysis of previous findings is necessary to correctly justify the author's own methods. In other cases, authors complete their discussion of one set of experiments and then move on to introducing a further set of methods.

Abstract Level Patterns

The results of analysing the *KT* values of the first and last event in each abstract are summarised in Table 1. Mostly, authors begin by stating known *Fact*s as a scene-setting device for introducing their own work. The use of *KT* categories other than *Fact* at the start of abstracts is considerably less frequent, with *Analysis* and *Observation* as the next most common categories. Analysis of the *Source* dimension of these event
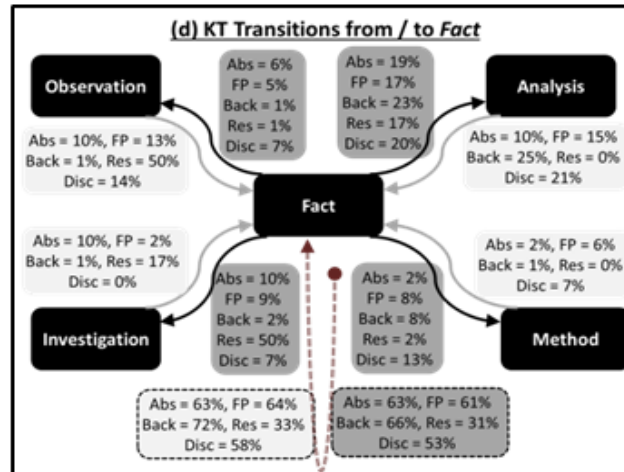
Fig. 3, continued. Continued to the next page.

types reveals that they often pertain to previous studies, indicating that a discussion of previous findings is also a common way to start.

Sometimes, scene-setting steps are omitted altogether, and the abstract launches directly into an explanation of the investigation to be undertaken. In rare cases, even the subject of investigation is missing, and the abstract starts by explaining the experimental setup and methodology. In the vast majority of cases, authors end their abstracts with an *Analysis*, presenting a summary or interpretation of their most important findings. However, there is a significant proportion of cases (15%) in which the abstract ends with an *Observation*. This can happen the when a significant experimental observation has occurred during the current study. Very occasionally, the abstracts end by presenting an investigative topic or method identified for further exploration.

Table 1. Relative frequencies of abstracts starting and ending with each *KT* category

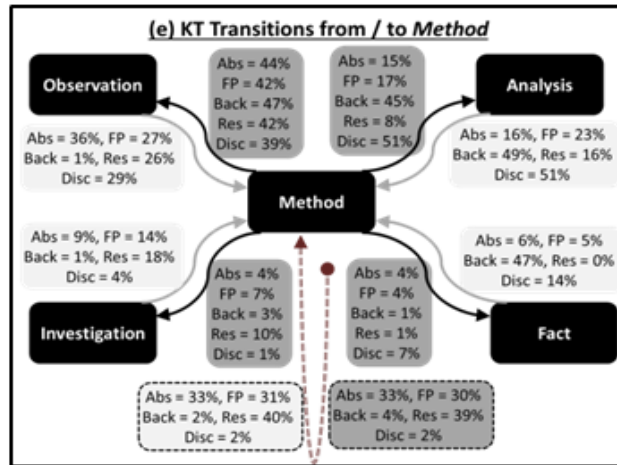| KT Category	Abstracts Starting With	Abstracts Ending With
Observation	10%	15%
Analysis	23%	78%
Investigation	9%	4%
Fact	54%	1%
Method	4%	2%

Fig. 3, continued.

Table 2 shows the most frequent extended transition patterns of KT values. 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.

Table 2. Key transition patterns for *KT* values in abstracts and their frequencies

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%

A significant number of abstracts follow a slightly different *KT* 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 [29], 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 due to partly to the specialised nature of many biomedical journals.

3.2   Certainty Level (CL)

Pair-wise Transitions

Figure 4 summarises the pair-wise transitions **from** and **to** adjacent events in the GENIA-MK corpus, according to the *CL* category assigned to them.

L3: This is a highly reflexive category, partly due to its high frequency of occurrence (92% of events in the GENIA-MK corpus). In terms of non-reflexive transitions, 6% of transitions from *L3* lead to *L2*, and only 1% to *L1*. 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.

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 avoid long chains of speculated events. This explain why significant proportion (40%) of transitions from *L2* lead back to *L3*. Interestingly, 6% of transitions from *L2* lead to *L1*. These are mostly the cases where slightly hedged analyses are followed by bolder (highly speculative) extensions and corollaries.

L1: For similar reasons as *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
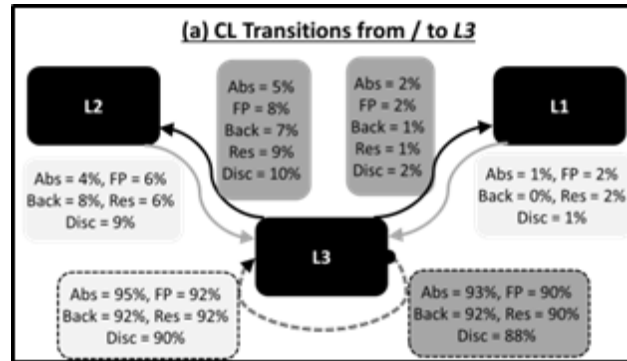
Fig. 4. Transitions from / to *CL* categories for Abstracts (Abs), Full Papers (FP), and the sections within full papers, i.e., Background (Back), Results (Res), and Discussion (Disc). Continued to the next page.

(less than 1% of all events) that they only account for around 1% of all transitions to *L3* and *L2*.

Abstract Level Patterns

The *CL* values of the first and last event in each abstract in the GENIA-MK corpus are summarised in Table 3. Almost all abstracts start with known facts, previous observations, analyses, or investigations, i.e., events expressed with absolute certainty of occurrence (*L3*). Although most abstracts end with analyses, authors will usually aim to have maximum impact at the end of their abstract, so as to encourage reading of the full text.

This means that where possible, hedging will either be absent, or only subtly expressed. A smaller, but still important percentage of terminal events are marked as highly speculative, sine impact can also be achieved by presenting analyses that are both highly speculative and highly innovative or controversial.

Table 3. Relative frequencies of abstracts starting and ending with different *CL* categories

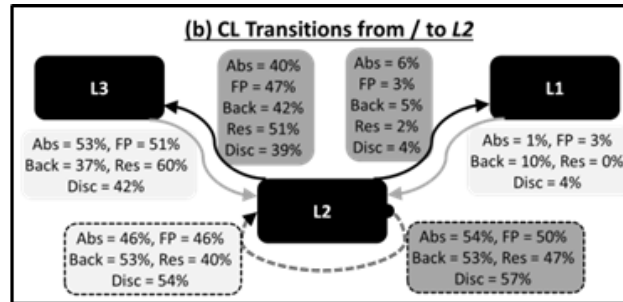| CL Category	Abstracts Starting With	Abstracts Ending With
L1	0%	19%
L2	1%	36%
L3	99%	45%

Fig. 4, continued. Continued to the next page.

Speculated events are completely absent in 28% of abstracts, 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 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. A small proportion of abstracts (5%) contain the 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, a significant proportion of abstracts (14%) contain the transition pattern *L3 → L1*, i.e., observations and confident analyses are followed directly by highly speculated analyses or hypotheses.

4 Full Papers

In this section we present a brief analysis of the meta-knowledge transitions observed in the *Background*, *Results*, and *Discussion* sections of the FP-MK corpus.

4.1   Knowledge Type (KT)

Figure 3 shows the summary of pair-wise transitions **from** and **to** adjacent events in the FP-MK corpus, according to *KT* categories. It in-
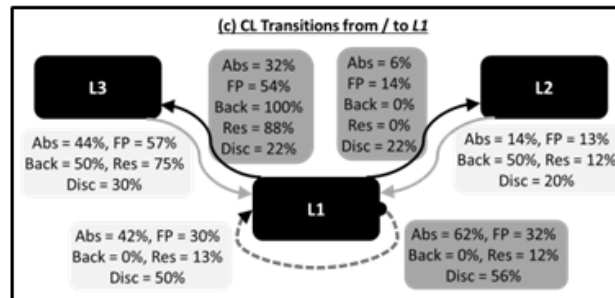
Fig. 4, continued.

cludes separate statistics for each of the main sections, as well as for the full papers as a whole.

Observation: Overall distributions of transitions from and to *Observation* in full papers are similar to those in abstracts. However, the reflexivity of *Observation* is slightly lower in full papers. This is partly because of the significantly higher proportion of transitions between *Observation* and *Analysis* in full papers. 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. 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*, 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, as evidenced by the much higher number of transitions from *Analysis* to *Observation* in full papers. 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 sec-

tions 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, due to a lower relative frequency of *Investigation* events in full papers. Full papers contain significantly higher numbers of transitions 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 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, which excludes many method descriptions from event annotation.

4.2   *Certainty Level (CL)*

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 due to more detailed analytical discussion often found in full papers. Moreover, unlike in abstracts, where the main aim is to try to sell the research results, the body of the paper provides greater opportunity for analysis and discussion. The percentage of *L3* to *L2* transitions is highest in the *Results* sections of the full papers. Authors may be confident about

some of their results, but not so confident about others. Fewer such transitions are found 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. If too many bold or controversial statements are made in the body of the paper, 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, they are more thinly spread out. The greater the number of highly speculative events that occur in sequence, the more wary the reader is likely to become.

5   Conclusion

In this paper, 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. We used the GENIA-MK corpus of abstracts and the FP-MK corpus of full paper to conduct our analyses. We examined a number of different types of discourse patterns, including patterns of pairwise transitions between events, considering *KT* and *CL* separately. Comparison of the results obtained for abstracts and full papers reveal that there are a number of subtle and significant differences in the patterns of local discourse-level shifts. For abstracts, we additionally considered extended transition paths. Whilst there are some clear patterns of *KT* and *CL* transitions in abstracts, these are by no means standard. Furthermore, while most abstracts follow a generic model of rhetoric/information moves, authors often skip certain moves, assuming that the reader is already familiar with the context.

As future work, we intend to broaden the scope of our study to incorporate different types of events and additional meta-knowledge dimensions across different domains. We also plan to investigate transition patterns within each section of full papers. Furthermore, with the

help of the BioDRB corpus, we intend to investigate 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.

References

1. 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.

2. Oda, K., Kim, J.-D., Ohta, T., Okanohara, D., Matsuzaki, T., Tateisi, Y., Tsujii, J.i.: New challenges for text mining: mapping between text and manually curated pathways. BMC Bioinformatics 9, 2008, S5.

3. 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.

4. Teufel, S., Carletta, J., Moens, M.: An annotation scheme for discourse-level argumentation in research articles. Proceedings of EACL, 1999, 110–117.

5. 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.

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

7. Marcu, D., Echihabi, A.: An unsupervised approach to recognizing discourse relations. Proceedings of ACL. Association for Computational Linguistics, 2002, 368–375.

8. Swales, J.: Genre Analysis: English in Academic and Research Settings: Cambridge Applied Linguistics. Cambridge University Press, 1990.

9. Teufel, S.: Argumentative Zoning. University of Edinburgh, 1999.

10. Marcus, M.P., Santorini, B., Marcinkiewicz, M.A.: Building a large annotated corpus of English: The Penn Treebank. Computational Linguistics 19, 1994, 313–330.

11. 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.

12. Mann, W.C., Thompson, S.A.: Rhetorical structure theory: Toward a functional theory of text organization. Text 8, 1988, 243–281.

13. 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.

14. Prasad, R., McRoy, S., Frid, N., Joshi, A., Yu, H.: The biomedical discourse relation bank. BMC Bioinformatics 12, 2011, 188.

15. 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.

16. Kim, J.-D., Ohta, T., Tsujii, J.: Corpus annotation for mining biomedical events from literature. BMC Bioinformatics 9, 2008.

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

18. Pyysalo, S., Ohta, T., Miwa, M., Cho, H.-C., Tsujii, J.i., Ananiadou, S.: Event extraction across multiple levels of biological organization. Bioinformatics 28, 2012, i575–i581.

19. Kim, J.-D., Pyysalo, S., Nedellec, C., Ananiadou, S., Tsujii, J. (eds.): Selected Articles from the BioNLP Shared Task 2011, Vol. 13. BMC Bioinformatics, 2012.

20. Nawaz, R., Thompson, P., McNaught, J., Ananiadou, S.: Meta-Knowledge Annotation of Bio-Events. Proceedings of LREC 2010, 2010, 2498–2507.

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

22. 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.

23. Miwa, M., Thompson, P., McNaught, J., Kell, D.B., Ananiadou, S.: Extracting semantically enriched events from biomedical literature. BMC Bioinformatics 13, 2012, 108.

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

25. 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 Scholarly Discourse (DSSD), 2012, 37–46.

26. Sauri, R., Pustejovsky, J.: FactBank: A Corpus Annotated with Event Factuality. Language Resources and Evaluation 43, 2009, 227-268.

27. Palmer, M., Gildea, D., Kingsbury, P.: The Proposition Bank: An Annotated Corpus of Semantic Roles. Computational Linguistics 31, 2005, 71–106.

28. Ruppenhofer, J., Ellsworth, M., Petruck, M., Johnson, C., Scheffczyk, J.: FrameNet II: Extended Theory and Practice, 2010.
29. Swales, J.: Genre Analysis: English in Academic and Research Settings. Cambridge University Press, 1990.

RAHEEL NAWAZ
NATIONAL CENTRE FOR TEXT MINING,
MANCHESTER INTERDISCIPLINARY BIOCENTRE,
UNIVERSITY OF MANCHESTER,
131 PRINCESS STREET, MANCHESTER, M1 7DN, UK
E-MAIL: <RAHEEL.NAWAZ@CS.MAN.AC.UK>

PAUL THOMPSON
NATIONAL CENTRE FOR TEXT MINING,
MANCHESTER INTERDISCIPLINARY BIOCENTRE,
UNIVERSITY OF MANCHESTER,
131 PRINCESS STREET, MANCHESTER, M1 7DN, UK
E-MAIL: <PAUL.THOMPSON@MANCHESTER.AC.UK>

SOPHIA ANANIADOU
NATIONAL CENTRE FOR TEXT MINING,
MANCHESTER INTERDISCIPLINARY BIOCENTRE,
UNIVERSITY OF MANCHESTER,
131 PRINCESS STREET, MANCHESTER, M1 7DN, UK
E-MAIL: <SOPHIA.ANANIADOU@MANCHESTER.AC.UK>