An Investigation on the Influence of Genres and Textual Organisation on the Use of Discourse Relations

Félix-Hervé Bachand, Elnaz Davoodi, and Leila Kosseim

Department of Computer Science & Software Engineering
Concordia University
Montreal, Canada
{f_bachan,e_davoo,kosseim}@encs.concordia.ca

Abstract. In this paper, we investigate some of the problems associated with the automatic extraction of discourse relations. In particular, we study the influence of communicative goals encoded in a given genre against another, and between the various communicative goals encoded between sections of documents of a same genre. Some investigations have been made in the past in order to identify the differences seen across either genres or textual organization, but none have made a thorough statistical analysis of these differences across currently available annotated corpora. In this paper, we show that both the communicative goal of a given genre and, to a lesser extend, that of a particular topic tackled by that genre, do in fact influence in the distribution of discourse relations. Using a statistically grounded approach, we show that certain discourse relations are more likely to appear within given genres and subsequently within sections within a genre. In particular, we observed that *Attributions* are common in the newspaper articles genre while *Joint* relations are comparatively more frequent in online reviews. We also notice that *Temporal* relations are statically more common in the methodology sections of scientific research documents than in the rest of the text. These results are important as they give clues to allow the tailoring of current discourse taggers to specific textual genres.

1 Introduction

Consider the simple discourse: *Writing a scientific paper takes time; we wrote this one in two months.* In a coherent text, textual units are not understood in isolation but in relation with each other through discourse relations that may or may not be explicitly marked. The fact that "we" wrote this paper in two months, *illustrates* that writing a scientific paper takes time. Research on discourse analysis tries to model the coherence relations that hold between textual units, and these allow us to interpret the text and understand the communicative purpose of its units. This, in turn, is useful for many Natural Language Processing (NLP) applications such as automatic summarisation, question answering and text simplification. The objective of this paper is to explore the relationship between genre and textual organisation and the use of discourse relations.

A. Gelbukh (Ed.): CICLing 2014, Part I, LNCS 8403, pp. 454–468, 2014.

The task of automatic discourse relation extraction is a particularly difficult one. One important difficulty stems from the need for the system to be aware of the rhetorical purpose of the discourse on several levels. The rhetorical structure of a document can be divided into several levels of abstraction, from the general, down to the more specific. Discourse parsers available today (eg. [1] [2] [3]) attempt to extract rhetorical relations between Elementary Discourse Units (EDUs) without trying to build to the highest level of discourse relations schemas, namely the textual genre. For our purpose, we consider three levels of abstractions related to rhetorical structures: the genre, the sections, and the relations between individual EDUs. We argue that in order to extract discourse relations effectively, a system should consider the higher level rhetorical structures that we describe here as genre and section. By genre we mean that texts can have a variety of communicative goals [4]. Examples of genres include: instructional texts, reviews, scientific papers, newspaper articles, etc. At a lower level, we consider the textual organization of the document. By that we mean that the documents could be separated into different sections and subsections, each emphasising a lower-level communicative purpose. For example, given a scientific paper, the sections will typically include: abstract, introduction, methodology, results, etc. It should be noted that different genres will typically exhibit different textual organisations. Compare our previous example of a scientific paper to a review of a film. The likelihood of the appearance of a methodology section in such a document is very low. Instead we are expecting sections such as plot description, criticism, conclusion, etc. This shows that the higher level genre distinction can be used to better identify the more fine grained textual organisation categorizations. This is in line with the hierarchical view of discourse analysis presented in [5]. Based on this, it seems intuitive that the distribution of the various types of discourse relations be influenced by its occurrence in a given section of the document as opposed to another. Both the genre and, subsequently, the textual organization are important features to be considered in the automatic tagging of discourse relations.

2 Previous Work

Currently available discourse parsers do not take genre and textual organisation into consideration when extracting discourse relations (eg. [1] [2] [3]). To some extend, [2] and [3] estimate the influence of textual organisation by using the distance between a relation and the beginning of the text. Neither, however, take the genres or sections of texts into account in a definitive sense.

A few attempts have already been made at investigating the relation between genres and textual organisation and their influence on the distribution of discourse relations.

Bonnie Webber's investigation [6] shows that genre does in fact appear to play a role in the distribution of discourse relations. In order to reach this conclusion, the author performed a frequency analysis of the PDTB corpus [7]. She split the corpus into four distinct sections, each identifying a specific genre. The *news*

section accounts for the largest portion of the corpus, with 1902 documents. The remaining 208 documents are split into *essays*, *summaries*, and *letters*. The author observed that in the case of labelling implicit relations, which are those that are not marked explicitly by expressions such as *therefore* and *in order to*, especially when such relations appear in between sentences, the genre appears to be a worthwhile feature to investigate. Another interesting point relates the overall structure of discourse relations across a given document. For example, a news article might start by giving an effective summary of its contents, while an essay is less likely to do so. This leads to the hypothesis that we should not only consider the distribution of relations one at a time, but we should also consider sequences of such relations and the influence of genre on the observed patterns. This is similar to the notion of rhetorical schemas described in [5].

Another interesting research deals with the concept of "stages" [8]. These are similar in nature to our notion of textual organisation. The author studied a corpus of movie reviews written by non-professionals and aggregated from various web-sites such as *Rotten Tomatoes* and *Epinions*, and found that such reviews are typically organized in five sections: *subject matter*, *plot description*, *character descriptions*, *background* and *evaluation*. These sections could be segmented in two larger communicative goals: description and evaluation, that usually appear in this order. In addition, [8] observed that the evaluation sections tend to contain more evaluative and subjective words. On the other hand, descriptive sections tend to contain more temporal connectives, as well as more causal-type connectives. These observations are relevant to our purpose as the appearance of such connectives hints towards the existence of certain discourse relations.

Another interesting work is that of [9] which argues that discourse relations can be used as a feature to segment a given document on various topics. This shows that there does exist a relation between discourse relations and document sections. In order to evaluate their hypothesis, the authors constructed a corpus of 140 texts in Brazilian Portuguese picked from a number of sections of mainstream news agencies. These were manually annotated using the Rhetorical Structure Theory (RST) framework [5] and split into various topics [10] (or what we refer to as sections). Their conclusion is that some relations tend to be more frequent around topic boundaries, while others were never recorded to occur around these same boundaries. The authors evaluated topic segmentation based on RST type annotation and noticed an improvement over their baseline implementation. This again appears to show that a relation between the distribution of discourse relations and sections does in fact exist.

However, to our knowledge, no previous work has attempted to measure empirically the influence of genre and textual organisation on the distribution of discourse relations using today's large scale annotated discourse corpora.

3 Methodology

In order to measure the influence of genre and textual organisation in the task of automatic discourse relation extraction, we analysed the distribution of discourse relations across various corpora spanning over various genres. Within

some of these corpora, we also identified sections and analysed the distribution of discourse relations across these sections.

3.1 Corpora

The surge of interest in computational discourse analysis could not have happened without the availability of large-scale annotated corpora. The first major effort was the RST Discourse TreeBank (RST-DT) [11], which uses then followed by: Graphbank [12], the Discourse Relations Reference Corpus (DRRC) [13] and the Penn Discourse TreeBank (PDTB) [7] which included 3165 documents (50,000 sentences) tagged using [14]'s model. Through strict annotation guidelines, these resources attained a high inter-annotator agreement, which made them usable for training machine learning techniques. In addition, the field of BioNLP became interested in the extraction of the *causality* relation in bio-medical texts and developed the BioCause corpus [15] and their own shared-tasks. In 2011, [16] took a larger view of the problem by tagging all relations in bio-medical texts and developed the Biomedical Discourse Relation Bank (BioDRB).

Most work on computational discourse analysis are based on two principal frameworks for the annotations of discourse level relations. The first, Rhetorical Structure Theory (RST) [5], was conceived in order to fill the need for a framework that could be used in tasks related to natural language generation. A detailed set of annotation guidelines based on the RST framework was later created by Marcu, et al. [17]. More recently, the Penn Discourse Tree Bank [7], makes use of a new annotation framework for discourse structures. Guidelines for the purpose of creating corpora within this framework were penned by [14]. PDTB has now become one of the most widely used corpora due to its size and annotation that attempt to remain framework agnostic.

For the purpose of our work, we used the following corpora:

RST Discourse Treebank. The RST Discourse treebank [11] consists of 385 articles from *The Wall Street Journal* annotated on discourse relations, with over 20,000 EDUs which are related together and tagged with RST's 78 relations. Given the source of these documents, they are generally written in a formal language.

Maite Taboada's Review Corpus. A second corpus we considered which also uses the RST framework, as well as Marcu's annotation guidelines is Taboada's Review corpus [18] [19]. This corpus is composed of 400 reviews gathered from the *Epinions* website, with over 12,000 discourse relations identified. These reviews are authored by non-professionals. As a result, the type of language used in these documents tend to be more informal and the overall structures seem to be somewhat more liberal.

Penn Discourse Tree Bank. The Penn Discourse Tree Bank (PDTB) is a large scale corpus which, much like RST, annotates discourse level relations [7].

The PDTB annotation style is an attempt at providing annotations which are theory-neutral. The annotation guidelines [14] were used in the creation of a number of corpora. They describe 43 discourse relations, some of which are hierarchically related. The original corpus covers the entire *Wall Street Journal* section of the Penn Treebank. The corpus is composed of 2304 texts, which are marked with over 40,000 relations.

Biomedical Discourse Relation Bank. The last corpus we investigated in our work is the Biomedical Discourse Relation Bank (BioDRB) [16]. It is composed of 24 open-access research papers in the biomedical field. Nearly 6000 relations were marked using the PDTB framework. An interesting feature of this corpus is that each document is split into several sections. These sections can be used as the basis of our investigation on textual organization (see Section 4.2)

3.2 Inter-framework Discourse Mapping

Since the annotation guidelines used for the corpora we used ([8] [14] [16] [20]) differ in some manners, some work had to be performed in order to map the different discourse relations across corpora. The RST-DT contains 78 discourse relations, grouped into 18 meta-relations, while the PDTB is built from 43 discourse relations which can be grouped overall into 4 broad categories. Although both the RST-DT [11] and the online reviews [19] [18] are based on the RST framework, they do not use exactly the same set of relations. Mapping between the RST Discourse Treebank and the Online review corpus was performed by only considering the *meta-relations* of each. Since the major differences in annotations between RST-DT and the reviews corpus are found in the finner grained relation types, using these higher level *meta-relations* allowed us to perform a sensible mapping between annotations. For example, both RST-DT and Taboada's Review Corpus include relations that can be grouped under the *Contrast* relation, even if some of these exact relations differ in both corpora. Similarly, although both the PDTB and the BioDRB corpora are grounded on the PDTB annotation guidelines [14], the BioDRB annotations differ in a number of ways. Some modifications were made by the authors of this corpus as they found that certain aspects of the original framework were inappropriate for their task. [16] provides a mapping between their new relations set. Because of this, we converted the data from the original PDTB corpus into the new relations of the BioDRB corpus, following the descriptions provided. In order to adequately compare the PDTB and BioDRB corpora, we relied on the descriptions given in [16] which detail the changes made to the original PDTB guidelines in order to obtain those used in the creation of the more recent corpus. Since this description shows how the authors converted the original PDTB annotation guidelines to the ones used in the creation of the bio-medical documents corpus, we have followed the same path and used the relations described in [16] while comparing these two corpora. Details on how to map the original PDTB relations to those we used are given in this same paper.

3.3 Log Likelihood Ratio

In order to identify statistically significant differences between genres and textual organization, we performed frequency profiling using the log likelihood ratio described in [21]. This measure allows us to compare the distribution of discourse relations across multiple corpora and sort them according to the importance of their relative frequencies. It then allows to identify the most relevant data points, but qualitative examination must subsequently be performed. The resulting numbers themselves only provide a measure of which discourse relations are statistically more informative. As described in [21] the log likelihood ratio for a given relation between corpora a and b is computed as:

$$LL = 2 \times \left(\left(O_a \times log \left(\frac{O_a}{E_a} \right) \right) + \left(O_b \times log \left(\frac{O_b}{E_b} \right) \right) \right) \tag{1}$$

where:

$$E_a = \frac{N_a \times (O_a + O_b)}{N_a + N_b}, E_b = \frac{N_b \times (O_b + O_a)}{N_b + N_a} \tag{2}$$

N_i corresponds to the total count of all relations in a given corpus, and O_i corresponds to the count of the relation for which we are currently making calculations in that corpus. The second and third formulas gives us the *expected values*, which are then used, as E_i, in the first formula.

4 Analysis

4.1 Distributions of Discourse Relations across Genre

We first studied the influence of genre on the distribution of discourse relations. To do so, we split the corpora described in Section 3.1 into two categories: RST framework corpora and PDTB framework corpora.

RST Framework Corpora. The first two corpora we analysed both use the RST framework [5] and guidelines [17]: RST-DT and the Taboada Review Corpus (see Section 3.1). From the RST-DT corpus, we only selected the documents that have been identified as newspaper articles in [6], leaving out "erratas", "letters", and "summaries", in order to limit our investigations to documents that are news stories. On top of the genres themselves being different, it should be noted that the newspaper articles of the RST-DT corpus are written in a very formal language, while the online reviews of the Taboada's corpus tend to be much more informal, as indicated in Section 3.3. In order to view the differences in terms of discourse relations between these two corpora, we calculated the log-likelihood ratio [21]. These results are shown in Table 1.

Table 1 shows the relations which appear to vary in a statistically significant manner while comparing the two corpora. The most obvious statistical differences stem from *Joint*, *Attribution*, *Enablement*, *Same-Unit*, *Background*, and *Elaboration*.

Table 1. Log Likelihood Ratio between RST-DT and Taboada's Reviews Corpus Using RST's Meta-Relations

Relation	RST-DT	Reviews	LL ratio
Joint	10.55%	33.35%	1,637.57
Attribution	11.07%	0.01%	1,546.38
Enablement	18.31%	3.02%	1,321.68
Same-Unit	9.36%	0.01%	1,305.96
Background	2.99%	13.23%	940.39
Elaboration	25.79%	9.25%	915.17
Contrast	4.90%	16.59%	888.74
Cause	2.55%	9.59%	578.64
Condition	1.04%	5.92%	517.30
Comparison	1.49%	0.01%	199.42
Explanation	3.41%	1.13%	134.44
Topic-Change	1.01%	0.01%	132.42
Textual-organization	0.88%	0.01%	114.62
Temporal	2.37%	4.73%	101.87
Summary	0.71%	0.14%	45.14
Topic-Comment	0.84%	0.26%	35.03
Manner-Means	0.73%	0.54%	3.41
Evaluation	1.97%	2.17%	1.25

A number of observations can be made from Table 1. First, *Elaborations* appear much more frequently in newspaper articles than in online reviews (25.79% vs. 9.25%). In fact, this single relation accounts for a quarter of the relations of the first corpus. It does not seem surprising that the *Elaboration* relation be used so often in news paper articles, as we would expect texts to bring forward an idea and then elaborate on it. This type of pattern can probably be expected regardless of genre. What we notice, however, is that while *Elaboration* is frequent in newspaper articles, the *Joint* relations are roughly three times more likely in the online review corpus (10.55% vs. 33.35%). The *Joint* relation is described in the annotation guidelines [17] as a pseudo-relation which should be used by annotators when no other relation seems appropriate. What these two distributions seem to indicate has more to do with the fact that the review corpus is written less formally than the newspaper articles of the *The Wall Street Journal*. What we mean by that is that, not only is the task of identifying discourse relations a difficult one, but using such relations appropriately is also difficult. The flow of a discourse redacted by a professional writer and with the help of an editorial staff is likely to be more easily identifiable than that of an amateur reviewer posting online, without any sort of peer review. For these reasons, we believe that the occurrence of such an elevated number of the *Joint* relation in the review corpus is likely due to the differences in the writers capacity to make an adequate use of discourse relations and the inherent ability at writing in a coherent manner. A similar conclusion can be reached when observing the distribution of *Same-Unit* relations, more frequent in the RST-DT corpus (9.36% vs. 0.01%). This particular pseudo-relation is intended to represent embedded

relations. For example, consider this excerpt from `wsj_1362` where the EDUs marked are related as *Same-Unit*, while the first EDU contains a *Enablement* relation.

> [a reserve it is establishing to cover expected pollution cleanup costs at an Ohio plant][reduced its third-quarter net income by $1.9 million.]

The use of such embedded discourse relations can once again be attributed to better stylistic choices made by authors with a better grasp of the language used. *Attribution* relations are much more frequent in the newspaper corpus than in the review corpus (11.07% vs. 0.01%). This comes as no surprise as reported news should include a number of statements that are later attributed to their authors. On the other hand, this is not the type of discourse structure we would expect from a review. The *Enablement* relation serves to provide a description of a condition which enables a subsequent occurrence. Finding this relation type in a corpus of newspaper articles is not surprising as events are often described in order to introduce more recent occurrences. Examples of *Background* relations are more frequent in the Online Reviews corpus. This again seems logical, as we would expect observations of reviewers to be justified by providing background information. Such relations are therefore useful in the case of reviews.

PDTB Framework Corpora. We now turn our attention to corpora making use of the PDTB framework for the annotation of discourse relations. We compare here the Penn Discourse Tree Bank to the Biomedical Discourse Relation Bank (BioDRB).

Table 2. Log Likelihood Ratio between PDTB and Bio-medical corpora, using PDTB Relations

Relation	**PDTB**	**BioDRB**	**LL ratio**
Circumstance	0.05%	4.09%	354.86
Contrast	16.00%	5.01%	228.97
Background	0.29%	2.36%	104.12
Condition	3.45%	0.39%	99.42
Purpose	6.05%	10.96%	66.73
Instantiation	4.28%	1.56%	50.45
Concession	3.62%	5.85%	24.18
Temporal	10.75%	13.86%	17.43
Restatement	6.98%	9.47%	16.89
Conjunction	23.03%	18.92%	16.80
Alternative	1.17%	0.66%	5.87
Continuation	13.48%	15.19%	4.42
Exception	0.04%	0.16%	4.20
Reinforcement	1.27%	1.79%	4.01
Similarity	0.14%	0.09%	0.51
Cause	9.40%	9.63%	0.12

Table 2 shows the differences in the distribution of discourse relations between our two corpora. Once again, we see a number of statistical differences between our corpora. The most noticeable being *Circumstances*, *Contrast*, *Background*, *Condition*, and *Purpose*. First, *Circumstance* relations are favored in the BioDRB corpus. These relations are often used in order to explain the specific conditions of a given experiment and subsequently describe the observed results within said conditions. The *Contrast* relation appears to be favored within the PDTB corpus (16.00% vs. 5.01%). A common way of using such a relation when dealing with text of the newspaper article genre is to compare divergent opinions. For example, consider the following from `wsj_0047`:

> [A majority of an NIH-appointed panel recommended late last year that the research continue under carefully controlled conditions] [but the issue became embroiled in politics as anti-abortion groups continued to oppose federal funding]

This type of discourse is quite common when dealing with news items in the social sphere, as divergence of opinions is generally what make the news. On the other hand, the *Background* relations are more frequent in the BioDRB corpus (0.29% vs. 2.36%). Such relations are used when background information is provided in order to allow the reader to fully grasp the arguments being made. It is not surprising to find such a relation in scientific papers where claims are often made based on background knowledge or previous work. *Condition* relations appear in the PDTB corpus at a greater frequency (3.45% vs. 0.39%). These relations are often used to put forward conditions necessary for certain predictions to become reality. Such conditions do not need to be realized, they can simply be hypothetical. This is not an uncommon strategy for journalists, as exemplified in `wsj_0664`:

> [If the exchange falters in these moves,][it might once again fall behind its chief New York competitor, the Commodity Exchange.]

Finally, we notice that the distributions of relations of *Cause* and *Similarity* seem to be constant across the two textual genres studied.

4.2 Distributions of Discourse Relations across Sections

We now turn our attention to the influence of textual organisation on the distribution of discourse relations. Our hypothesis is that, just like with the higher-level communicative goal of the textual genre, the organisation of discourse into sections play a role in influencing the discourse relations employed in the lower-level communicative goals expressed though sections. For example, we believe that the distribution of discourse relations encountered in the *abstract* section of scientific papers should differ from those found in the *methodology* section. In order to evaluate this claim, we once again analysed the BioDRB corpus which is already split into sections. Each of these section refers to the usual sections found in scientific papers: *introduction*, *methodologies*, *results*, *abstracts*, and *discussions*. In order to discover statistically significant data, we again computed

Table 3. Distributions of Discourse Relations Across Sections in BioDRB

	Overall	**Introduction**		**Methods**	
Relation	Distribution	Distribution	LL ratio	Distribution	LL ratio
Alternative	0.92%	0.79%	0.07	0.59%	0.88
Background	3.07%	3.94%	1.07	1.17%	13.84
Cause	12.32%	12.76%	0.16	2.49%	**89.83**
Circumstance	4.10%	1.89%	**11.24**	1.17%	23.33
Concession	7.11%	9.45%	5.31	0.59%	**77.58**
Condition	0.61%	0.31%	1.33	1.02%	1.94
Conjunction	18.81%	16.69%	2.98	14.79%	9.73
Continuation	11.20%	12.76%	2.55	17.86%	33.30
Contrast	6.63%	4.25%	6.36	1.46%	43.86
Exception	0.28%	0.31%	0.03	0.44%	0.63
Instantiation	2.01%	2.83%	2.71	0.15%	21.82
Purpose	11.96%	14.8%	4.83	12.45%	0.16
Reinforcement	2.37%	2.36%	0.01	0.59%	14.42
Restatement	8.44%	10.08%	1.94	7.03%	2.41
Similarity	0.25%	0.16%	0.33	0.15%	0.45
Temporal	9.92%	6.61%	8.46	38.07%	**478.55**

	Results		**Abstracts**		**Discussions**	
Relation	Distribution	LL ratio	Distribution	LL ratio	Distribution	LL ratio
Alternative	0.68%	0.75	0.33%	1.39	1.46%	5.42
Background	3.15%	0.04	5.35%	3.77	3.65%	0.77
Cause	11.5%	0.78	11.37%	0.21	19.07%	**53.27**
Circumstance	9.63%	**108.90**	4.68%	0.22	1.28%	37.70
Concession	6.3%	1.64	7.36%	0.02	10.68%	24.93
Condition	0.09%	10.62	0.67%	0.01	1.09%	5.06
Conjunction	23.17%	11.93	26.42%	**7.46**	17.88%	1.84
Continuation	9.63%	2.22	6.02%	**8.08**	7.85%	13.19
Contrast	10.14%	32.58	4.35%	2.50	7.48%	2.41
Exception	0.26%	0.05	0.33%	0.03	0.18%	0.60
Instantiation	0.85%	12.22	1.34%	0.70	3.92%	26.70
Purpose	12.35%	0.21	12.04%	0.00	9.58%	7.58
Reinforcement	1.28%	8.75	1.67%	0.64	4.65%	31.80
Restatement	10.65%	8.16	10.03%	0.77	6.02%	12.39
Similarity	0.26%	0.00	0.33%	0.07	0.36%	0.64
Temporal	0.09%	**260.37**	7.69%	1.51	4.84%	43.42

the log likelihood ratio [21]. This time, this measure was computed for each section with respect to the overall distribution of relations in the corpus.

The results shown in Table 3 show the relations that have a more statistically significant difference in distribution across sections. The most striking values of the log likelihood ratio are seen with *Temporal* relations which are significantly more frequent in the **Methods** (38.07% vs. 9.92%) and **Results** (0.09% vs. 9.92%) sections. This seems intuitive as we would expect a description of methodologies to include experimental steps to be taken in succession through time. Other values are also worth noting. The most statistically significant

difference observed in the **Introduction** section shows a slight tendency to disfavor *Circumstance* relations (1.89% vs. 4.10%). Such a relation is used to describe the conditions in which an event occurs, without the need for the event and the circumstances to influence each other. It seems relevant to use this relation in the **Results** section, where circumstances are first given in order to set the stage for the results observed. Such a discourse schema is not as useful in an introduction. The **Methods** section's second and third most significant distinctions are seen with the *Cause* (2.49% vs. 12.32%) and *Concession* (0.59% vs. 7.11%) relations. It should be noted that both these relations are less likely to appear in the **Methods** section. Again, the **Methods** tend to favor *Temporal* relations, describing successive steps in an experiment. The **Results** section shows *Temporal* (0.09% vs. 9.92%) and *Circumstance* (9.63% vs. 4.10%) relations to be the most significant differences. Seeing a higher frequency of *Circumstance* relations in this section seems again intuitive as the presentation of results are often made in the context of the circumstances observed during experimentation. In the **Abstracts** section, the *Conjunction* (26.42% vs. 18.81%) and *Continuation* (6.02% vs. 11.20%) relations are the most statistically different. *Conjunction* relations are used in order to link EDUs as part of a list. This seems appropriate, especially in an **Abstract**, where statements are compressed due to constraints on length. In the **Discussion** section, *Cause* (19.07% vs. 12.32%) is statistically the most different in its distribution. Again, this appears intuitive as the discussion should explain the causes for the observed results. Finally, relations such as *Alternative*, *Exception*, and *Similarity* do not seem to be used differently across the sections studied.

Overall, Table 3 shows that even through a small investigation of the distributions of discourse relations, sections do in fact appear to play an important role and that these differences can be justified fairly naturally. One interesting observation is that the result of the log-likelihood ratio calculations shown in Table 3 are generally much lower than those seen in Tables 1 and 2. This suggests that the communicative goal of the textual genre has a larger influence over these distributions than the communicative goal of given sections in documents of a same genre.

Since currently, only the BioDRB corpus is segmented into sections, in order to further evaluate our claims on the influence of textual organisation, we proceeded with the following investigation: using the RST-DT corpus [11], we clustered the discourse relations used according to how far within the document they occur. Specifically, we counted the number of discourse relations in each document, and separated them in five pseudo-sections, each containing 20% of the total discourse relations found in the document. For example, a document with 10 discourse relations would have its first two grouped together, followed by the next two, and so on. With this simple heuristic, and assuming that the documents of our corpus all share the same general pattern, as dictated by the genre, we were able to identify in which portion of the documents certain relations are more likely to occur and approximate the notion of textual organisation. Once again, the log likelihood ratio is calculated by comparing a given pseudo-section to all the overall distribution.

Table 4. Distributions of Discourse Relations Per Pseudo-Sections in RST-DT

	Overall	**0-20%**		**20-40%**	
Relation	Distribution	Distribution	LL ratio	Distribution	LL ratio
Attribution	12.00%	10.86%	2.65	11.09%	1.39
Background	3.64%	4.76%	**7.71**	3.81%	0.26
Cause	3.04%	3.49%	1.59	2.75%	0.57
Comparison	1.68%	1.60%	0.08	1.37%	1.26
Condition	1.29%	1.01%	1.45	1.42%	0.38
Contrast	5.94%	5.90%	0.01	5.66%	0.24
Elaboration	31.28%	30.45%	0.53	32.91%	2.63
Enablement	2.30%	2.50%	0.43	2.55%	0.72
Evaluation	2.33%	2.25%	0.06	2.12%	0.39
Explanation	3.88%	3.72%	0.17	4.06%	0.25
Joint	13.45%	11.00%	**11.29**	13.14%	0.08
Manner-means	0.88%	0.74%	0.50	0.88%	0.00
Same-unit	11.10%	11.88%	1.28	11.65%	0.86
Summary	0.84%	2.14%	**37.90**	0.65%	0.99
Temporal	2.97%	2.88%	0.06	2.79%	0.19
Textual-organization	1.23%	1.74%	4.50	0.72%	5.59
Topic-change	1.19%	1.78%	6.36	0.74%	4.30
Topic-comment	0.98%	1.28%	2.16	0.95%	0.01

	40-60%		**60-80%**		**80-100%**	
Relation	Distribution	LL ratio	Distribution	LL ratio	Distribution	LL ratio
Attribution	11.76%	0.04	13.14%	2.94	14.60%	1.39
Background	3.18%	1.27	3.40%	0.28	3.47%	2.92
Cause	2.73%	0.67	3.47%	1.59	3.11%	0.96
Comparison	1.92%	0.87	1.71%	0.03	1.98%	0.06
Condition	1.06%	0.92	1.31%	0.02	1.78%	1.56
Contrast	6.04%	0.08	6.04%	0.08	6.78%	0.00
Elaboration	31.87%	0.52	29.19%	2.69	35.74%	0.00
Enablement	2.30%	0.00	1.74%	3.33	2.68%	0.03
Evaluation	2.32%	0.00	2.55%	0.56	2.68%	0.00
Explanation	4.37%	1.62	4.15%	0.53	3.58%	4.29
Joint	13.91%	0.54	14.38%	1.84	16.48%	1.90
Manner-means	0.77%	0.30	0.97%	0.27	1.13%	0.38
Mame-unit	10.89%	0.04	10.23%	1.36	12.19%	0.48
Summary	0.52%	3.21	0.34%	**8.66**	0.65%	2.66
Temporal	3.34%	1.18	3.27%	0.80	2.93%	1.64
Textual-organization	0.56%	**10.18**	1.13%	0.18	2.16%	**9.05**
Topic-change	0.92%	1.40	1.08%	0.20	1.56%	0.68
Topic-comment	0.81%	0.65	1.17%	0.96	0.79%	2.59

Table 4 shows the distribution of discourse relations, along with their log likelihood ratio, for each pseudo-section of the documents. A first observation is that a *Summary* is statistically more likely to occur at the onset of the document, and unlikely past half of the document, especially in the 60% to 80% pseudo-section. This seems to make sense given our newspaper article corpus, where documents are likely to start with a very brief summary of the news item, followed by the

detailed explanation. The *Background* relation is more frequently seen at the beginning of a document as well. This, again, seems intuitive as providing a background in order to contextualize a news item is a typical writing strategy. The *Joint* relation, on the other hand, is less likely to occur at the beginning of documents. We assume here that since the bulk of the information provided in such documents should be located towards the middle, it seems more likely that such a relation, which joins together said information, should occur in the body of a newspaper article. The *Textual-organization* relation, which links a sub-heading with its associated section, is noted to be unlikely towards the middle of such newspaper articles, but more likely towards the end. This is likely due to the use of sub-headings to introduce new items which are related to the news being covered in the article.

5 Conclusion and Future Work

In this paper, we have performed an analysis of the distributions of discourse relations across various genres and sections. Using currently available annotated corpora, which themselves use different discourse relation frameworks, we have studied how both genre and textual organisation affects the distribution of discourse relations at the unigram level. As the RST framework suggests, discourse analysis is a hierarchical process and the construction of a discourse starts at the top with the communicative goal of the textual genre, and subsequently trickles down to the sections and sub-sections and finally between individual EDUs. In particular, we observed that *Attributions* are much more common in newspaper articles than in online reviews, and that newspaper articles favor *Enablement* while online reviews favor *Joint* relations. *Circumstance* relations are favored in scientific papers compared to newspaper articles, while the opposite is true of the *Contrast* relation. Our investigation of lower-level communicative goals across sections shows that the *Temporal* relation is significantly different across sections, while a number of other relations provide significant statistical differences across specific sections, to a lesser degree. Our observations therefore suggest that a worthwhile approach to extracting discourse relations should take into account both the genre of the text, hopefully with access to annotated corpora within that same genre, and should then partition the text into sections which themselves provide an influence on the distributions of the low-level discourse relations. In addition, while the task of identifying discourse relations is difficult, the use of those same relations appropriately is difficult for the authors themselves. We believe that some of our results, especially when comparing newspaper articles to online reviews, hints towards how the use of a more formal language usually comes with a better use of discourse relations.

As future work, it would be interesting to study the distribution of specific sequences of discourse relations, by observing discourse bigrams and trigrams, as opposed to the distribution of unigrams alone. This would be a step towards the automatic creation of discourse schemas described within the RST framework. Future work also includes the analysis of discourse relations across other types

of textual genres such as poetry, political speech, but doing so is difficult to do objectively without properly annotated corpora.

Acknowledgement. The authors would like to thank the anonymous reviewers for their comments on an earlier version of the paper. This work was financially supported by an NSERC grant.

References

1. Soricut, R., Marcu, D.: Sentence level discourse parsing using syntactic and lexical information. In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, Edmonton, Canada, vol. 1, pp. 149–156 (2003)
2. Hilda: A discourse parser using support vector machine classification
3. Feng, V.W., Hirst, G.: Text-level discourse parsing with rich linguistic features. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, Jeju Island, Korea, vol. 1, pp. 60–68 (2012)
4. Swales, J.: Genre analysis: English in academic and research settings. Cambridge University Press (1990)
5. Mann, W.C., Thompson, S.A.: Rhetorical structure theory: A framework for the analysis of texts. IPRA Papers in Pragmatics 1, 79–105 (1987)
6. Webber, B.: Genre distinctions for discourse in the Penn Treebank. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, Suntec, Singapore, vol. 2, pp. 674–682 (2009)
7. Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A.K., Webber, B.L.: The Penn Discourse TreeBank 2.0. In: Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC), Marrakech, Morocco, pp. 2961–2968 (2008)
8. Taboada, M.: Stages in an online review genre. Text & Talk-An Interdisciplinary Journal of Language, Discourse & Communication Studies 31(2), 247–269 (2011)
9. Cardoso, P.C., Taboada, M., Pardo, T.A.: On the contribution of discourse structure to topic segmentation. In: Proceedings of the Special Interest Group on Discourse and Dialogue (SIGDIAL), Metz, France, pp. 92–96 (2013)
10. Cardoso, P.C., Maziero, E.G., Castro Jorge, M., Seno, E.M., Di Felippo, A., Rino, L.H., Nunes, M.: Cstnews-A discourse-annotated corpus for single and multi-document summarization of news texts in Brazilian Portuguese. In: Proceedings of the 3rd RST Brazilian Meeting, Brazil, pp. 88–105 (2011)
11. Carlson, L., Okurowski, M.E., Marcu, D.: RST Discourse Treebank. Linguistic Data Consortium, University of Pennsylvania (2002)
12. Wolf, F., Gibson, E., Fisher, A., Knight, M.: Discourse Graphbank. Linguistic Data Consortium, Philadelphia (2004)
13. Taboada, M., Renkema, J.: Discourse relations reference corpus. Simon Fraser University and Tilburg University (2008), `http://www.sfu.ca/rst/06tools/discourse_relations_corpus.html`
14. Prasad, R., Miltsakaki, E., Dinesh, N., Lee, A., Joshi, A., Robaldo, L., Webber, B.L.: The Penn Discourse Treebank 2.0 annotation manual. Technical Report, Institute for Research in Cognitive Science, University of Pennsylvania (2007), `http://www.seas.upenn.edu/pdtb/PDTBAPI/pdtb-annotation-manual.pdf`

15. Mihaila, C., Ohta, T., Pyysalo, S., Ananiadou, S., et al.: Biocause: Annotating and analysing causality in the biomedical domain. BMC Bioinformatics 14(2) (2013)
16. Prasad, R., McRoy, S., Frid, N., Joshi, A., Yu, H.: The biomedical discourse relation bank. BMC Bioinformatics 12, 188
17. Marcu, D.: Instructions for manually annotating the discourse structures of texts (1999), http://www.isi.edu/marcu
18. Taboada, M., Anthony, C., Voll, K.: Methods for creating semantic orientation dictionaries. In: Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC), Genova, Italy, pp. 427–432 (2006)
19. Taboada, M., Grieve, J.: Analyzing appraisal automatically. In: Proceedings of AAAI Spring Symposium on Exploring Attitude and Affect in Text, Stanford University, CA, pp. 158–161 (2004)
20. Carlson, L., Marcu, D., Okurowski, M.E.: Building a discourse-tagged corpus in the framework of rhetorical structure theory. In: Proceedings of the Second SIGdial Workshop on Discourse and Dialogue, SIGDIAL 2001, Aalborg, Denmark, vol. 16, pp. 1–10 (2001)
21. Rayson, P., Garside, R.: Comparing corpora using frequency profiling. In: Proceedings of the Workshop on Comparing Corpora, Hong Kong, pp. 1–6 (2000)