

# **An Approach to Corpus-based Discourse Analysis: The Move Analysis as Example**

**THOMAS A. UPTON AND MARY ANN COHEN**

## **Abstract**

This article presents a seven-step corpus-based approach to discourse analysis that starts with a detailed analysis of each individual text in a corpus that can then be generalized across all texts of a corpus, providing a description of typical patterns of discourse organization that hold for the entire corpus. This approach is applied specifically to a methodology that is used to analyze texts in terms of the functional/communicative structures that typically make up texts in a genre: move analysis. The resulting corpus-based approach for conducting a move analysis significantly enhances the value of this often used (and misused) methodology, while at the same time providing badly needed guidelines for a methodology that lacks them. A corpus of ‘birthmother letters’ is used to illustrate the approach.

Biber et al. (2007) explore how discourse structure and organization can be investigated using corpus analysis; they offer a structured, seven-step corpus-based approach to discourse analysis that results in generalizable descriptions of discourse structure. This article draws on the themes in this book, but focuses in particular on analyses that use theories on communicative or functional purposes of text as the starting point for understanding why texts in a corpus are structured the way they are, before moving to a closer examination and description of the linguistic characteristics and overall organizational tendencies reflective of the corpus. Biber et al. (2007) refer to this as a ‘top-down approach’ to the analysis of discourse structure. (In a bottom-up approach, the lexical and/ or form-focused corpus analysis comes first, and the discourse unit types emerge from the corpus patterns. See Biber et al., 2007, for discussion.) The primary goal of this article is to provide a description of the process for carrying out a corpus-based discourse study using the approach introduced by Biber et al. (2007), showing how it can be applied to move analyses.

## **1. Perspectives on discourse analysis**

### **1.1 THREE GENERAL CATEGORIES OF DISCOURSE ANALYSIS**

Schiffrin et al. (2001), in their introduction to *The Handbook of Discourse Analysis* (p. 1), note that definitions of ‘discourse analysis’ can be grouped into three general categories: 1) the study of language use; 2) the study of linguistic structure ‘beyond the sentence’; and 3) the study of social practices and ideological assumptions that are associated with language and/or communication.

The study of *language use* focuses on traditional linguistic constructs, such as phrase and clause structures, but generally addresses the problem of why languages have structural variants with nearly equivalent meanings (e.g. particle movement, as in ‘pick up the book’ versus ‘pick the book up’). The study of linguistic structure *beyond the sentence* focuses on a larger object of study: extended sequences of utterances or sentences, and how those ‘texts’ are constructed and organized in systematic ways. Although studies of this type are removed from the traditional concerns of structural linguistics (which focuses mostly on phrasal and clause syntax), the two share a primary focus on linguistic form and how language structures are used for communication.

In contrast, socio-cultural approaches to discourse focus either on the actions of participants in particular communication events or on the general characteristics of speech/discourse communities in relation to issues such as power and gender. These approaches typically are not concerned with understanding the linguistic forms used in those texts. This is not to say that these three different perspectives require isolated approaches to discourse analysis; Bhatia (1993b), among others, have underscored the need to investigate texts from multiple perspectives.

## **1.2 CORPUS-BASED INVESTIGATION OF DISCOURSE STRUCTURE**

Of the three general approaches to discourse analysis described above, only the first two focus on the linguistic characteristics of texts and discourse. The first focuses on the distribution and functions of surface linguistic features; these are typically corpus studies of language use in discourse. The second approach focuses on the internal organization of texts; that is, the discourse studies of linguistic structure beyond the sentence in particular texts. Studies from this perspective have usually been qualitative and based on detailed analyses of a small number (often fewer than five) of texts; these studies typically focus on the internal structure of a few texts from a single genre, such as scientific research articles.

Biber et al. (2007: 10) see the interface of these two perspectives as one of the current challenges of corpus linguistics:

Is it possible to merge the analytical goals and methods of corpus linguistics with those of discourse analysis that focuses on the structural organization of texts? Can a corpus be analyzed to identify the general patterns of discourse organization that are used to construct texts, and can individual texts be analyzed in terms of the general patterns that result from corpus analysis?

Few studies have attempted to combine these two research perspectives. On the one hand, most corpus-based studies have focused on the quantitative distribution of lexical and grammatical features, generally disregarding the language used in particular texts and higher-level discourse structure or other aspects of discourse organization. On the other hand, most qualitative discourse analyses have focused on the analysis of discourse patterns in a few texts from a single genre, but they have not provided tools for empirical analyses that can be applied on a large scale across a number of texts or genres. As a result, we know little at present about the general patterns of discourse organization across a large representative sample of texts from a genre.

The advantages of a corpus approach for the study of discourse, lexis, and grammatical variation include the emphasis on the representativeness of the text sample, and the computational tools for investigating distributional patterns across discourse contexts. One specific research emphasis for discourse studies of structure ‘beyond the sentence’ has been the attempt to segment a text into higher-level structural units. The ‘units of analysis’ in corpus-based studies of discourse structure must be well-defined discourse units: the segments of discourse that provide the building blocks of texts.

One of the major methodological problems to be solved by any corpus-based analysis of discourse structure, then, is deciding on a unit of analysis. That is, the first step in an analysis of discourse structure is to identify the internal discourse segments of a text, corresponding to distinct propositions, topics, or communicative functions; these discourse segments become the basic units of the subsequent discourse analysis. For a corpus study of discourse structure, all texts in the corpus must be analyzed for their component discourse units.

## **2. Identifying structural discourse units**

There are many types of analyses that have been conducted from the perspective of discourse analysis as the study of linguistic structure ‘beyond the sentence’, with the focus on communicative/functional organization (see Biber et al., 2007, for a review of these). However, in order to investigate higher-level discourse structure we need to start with theories of discourse that take into consideration the structure of whole texts. Theories describing the overall forms of discourse – what Van Dijk has called ‘schemata’ or ‘superstructures’ because they provide the ‘overall organizational pattern’ of different types of discourse (Van Dijk, 1988: 26) – have been developed for different types of texts such as exposition, argumentation, and narration. These include, among others, problem-solution patterns (Hoey, 1986) and story grammars (Mandler and Johnson, 1977). There are other approaches to the analysis of text structure that could be classified as being top-down in nature (see e.g. Mann and Thompson, 1992).

Connor (1996) has pointed out that the above kinds of analyses provided an important development in written discourse analysis. Researchers became keenly aware that different textual modes (e.g. narration, exposition, argumentation) use different discourse structures. Unlike the study of cohesion, for example, the analysis of superstructures is specific to a text type. The increased interest in specific genres has further stimulated research on discourse structures of texts.

‘Move analysis’ (Swales, 1981, 1990) is by far the most common example of such a specific genre-level analysis. Move analysis was developed as a top-down approach (where the focus is on meaning and ideas) to analyze the discourse structure of texts from a genre; the text is described as a sequence of ‘moves’, where each move represents a stretch of text serving a particular communicative (that is, semantic) function. The analysis begins with the development of an analytical framework, identifying and describing the move types that can occur in this genre: these are the functional/communicative distinctions that move types can serve in the target genre.

Subsequently, selected texts are segmented into moves, noting the move type of each move. The overall discourse structure of a text can be described in relation to the sequence of move types. For example, a research article might begin with a move that identifies the topic and reviews previous research, followed by a move that identifies a gap in previous research, followed by moves that outline the goals of the present study, summarize the major findings, and/or outlines the organization of the article.

Until recently, top-down approaches (including move analysis) have not been

applied to an entire corpus of texts because it is highly labor-intensive to apply a top-down analytical framework to a large corpus. However, this investment of labor pays off by enabling more detailed but generalizable analyses of discourse structure across a representative sample of texts from a genre. In particular, once a corpus of texts has been coded for moves, the typical linguistic (lexical and grammatical) characteristics of each move type can be analyzed, something which is rarely done for move analyses, allowing us to better understand the syntactic features of moves identified by their semantic roles. It is then possible to identify the sequences of move types that are typical for a genre, and against that background it is also possible to identify particular texts that use more innovative sequences of move types. The value of this sort of information is particularly important in instructional contexts, such as English for specific purposes classes or professional writing workshops.

### **3. An approach to conducting top-down corpus-based analyses of discourse organization**

To achieve generalizable corpus-based descriptions of discourse structure with communicative or functional purposes of text as the starting point of analysis, Biber et al. (2007) outline seven major analytical steps that must be followed, as shown in Table 1. This will be referred to in this article as the *Biber Connor Upton (BCU) Approach*. The seven steps of this approach will be described more fully in section 5 later.

In a top-down approach to discourse analysis, the first step is to develop the analytical framework, determining the set of possible discourse unit types based on an a priori determination of the major communicative functions that discourse units can serve in these texts. That framework is then applied to the analysis of all texts in the corpus. Thus, when texts are segmented into discourse units it is done by identifying a stretch of discourse of a particular type – that is, that serves a particular communicative function. Once these discourse units are identified, they are then analyzed and described for their lexical/grammatical features; complete texts and then the full corpus are then analyzed and described by organizational patterns.

TABLE 1. BCU Approach: *top-down corpus-based analyses of discourse organization*<sup>a</sup>

<i>Required step in the analysis</i>	<i>Realization in this approach</i>
1. Communicative/functional categories	Develop the analytical framework: determine set of possible functional types of discourse units, that is, the major communicative functions that discourse units can serve in corpus
2. Segmentation	Segment each text into discourse units (applying the analytical framework from Step 1)
3. Classification	Identify the functional type of each discourse unit in each text of the corpus (applying the analytical framework from Step 1)
4. Linguistic analysis of each unit	Analyze the lexical/grammatical characteristics of each discourse unit in each text of the corpus
5. Linguistic description of discourse categories	Describe the typical linguistic characteristics of each functional category, based on analysis of all discourse units of a particular functional type in the corpus
6. Text structure	Analyze complete texts as sequences of discourse units shifting among the different functional types
7. Discourse organizational tendencies	Describe the general patterns of discourse organization across all texts in the corpus

<sup>a</sup>Adapted from Biber et al. (2007).

## 4. Refining the move analysis

As noted above, the move analysis is one of the most common examples of a text-level analysis of discourse structure. While move analysis was originally developed as a tool to teach non-native speakers the rhetorical structures of research articles (Swales, 1981), this framework has been successfully extended to other areas of English for Specific Purposes (ESP) instruction. Swales's framework of move analysis has stimulated substantial research on the rhetorical structures of academic and professional texts. In academic writing, it has been applied to academic disciplines including biochemistry (Kanoksilapatham, 2005), biology (Samraj, 2002), computer science (Posteguillo, 1999), and medicine (Williams, 1999), as well as on a variety of academic genres, including university lectures (Thompson, 1994) and textbooks (Nwogu, 1991). More recently, professional discourse has also been examined through the lens of move analysis, including legal (Bhatia, 1993b) and philanthropic discourse – focusing on direct mail letters (Upton, 2002; Upton and Connor, 2001) and grant proposals (Connor and Upton, 2004).

Even so, with some exceptions (e.g. Kanoksilapatham, 2005), move analyses

have not been approached from a fully corpus-based perspective. In addition, the methods for conducting move analyses, especially as described in published studies, have all too frequently been vague and lacking standardization. The primary purpose for this article is to offer general but clear steps for doing a move analysis, relating the methodology to the steps required for a corpus-based analysis of discourse organization – the *BCU Approach* – introduced earlier. Examples from a move analysis study on ‘birthmother letters’ will be used to illustrate this approach. There are no clear guidelines in the literature for how to go about doing a corpus-based move analysis, and this article seeks to fill that gap. Our hope is that this discussion will provide a useful guide to novices in this fairly complex methodology and lead to better and more thorough analyses of genre move structures.

#### **4.1 BIRTHMOTHER LETTERS**

To illustrate the *BCU Approach* for conducting top-down corpus-based analyses of discourse organization, we will use examples from a study on birthmother letters. Due to the constraints of space, our description of the birthmother letters genre is limited, and is intended only to provide an example of the steps to the *BCU Approach*. (A more detailed description of this genre can be found in Cohen, 2007.)

Birthmother letters are the letters written by prospective adoptive parents to expectant mothers considering adoption plans for their unborn children. In many states, these letters are often the only legal way a couple hoping to adopt a child is allowed to make their intentions known to a birthmother. Even a casual reading of these letters shows that the wording in these letters is purposefully chosen and strategically placed by the writers. The birthmother letter study sought to understand what it is that makes one letter ‘successful’ and another not, based on the rhetorical moves used.

Birthmother letters are an interesting genre to study because they are very different from the academic (e.g. research articles) and professional (e.g. business letters) genres that have been typically analyzed. Bhatia (1993a) notes that a distinct genre has a specific communicative purpose, is identified and mutually understood by the community in which it is used, is usually highly structured, and is bound by ‘constraints’ that are readily noticed when broken. Birthmother letters fit this classification and are all the more interesting to study because it is a genre typically written by ‘lay people’ who are given only a brief orientation to the genre, not academics or professionals.

The corpus in this study consists of 46 birthmother letters written by couples who fit comparable profiles. The authors of these letters are all white, heterosexual, married couples and they are in a similar middle-class socioeconomic bracket. Twenty of the letters were each chosen by birthmothers after they were sent out once or no more than twice. The other 26 letters were sent out at least 15 times before they were selected. For the purpose of this project, the letters chosen after being sent out no more than once or twice will be referred to as ‘successful’ letters (S); the other letters will be referred to as ‘unsuccessful letters’ (U). The successful letters contain a total of 40,341 words. The average letter has a length of 2022 words, the longest at 4136 words and the shortest at 1094 words. The unsuccessful letters total 44,137 words. The average unsuccessful letter has a length of 1698 words, the longest at 2510 words and the shortest at 856 words. In the excerpts provided below, successful letters are each labeled S1 to S20; the unsuccessful letters are each labeled U1 to U26.

## **5. Overview of the move analysis in terms of the approach to conducting top-down corpus-based analyses of discourse organization**

As suggested earlier, the *BCU Approach* for doing top-down corpus-based analyses of discourse organization can be directly applied to move analyses, resulting in more thorough and informative results. Drawing on the framework introduced in Table 1, Table 2 outlines how these seven steps are realized in a move analysis. Each of these seven steps, as they apply to a corpus-based move analysis, is described below.

### **5.1 DEVELOP COMMUNICATIVE/FUNCTIONAL CATEGORIES**

The first step in the *BCU Approach* is to develop the analytical framework for analysis; that is, determine the set of communicative and functional categories that can be represented in the texts in the genre. In a move analysis, this first step is a two-part process. In order to identify the move categories for a genre, it is necessary to get a ‘big-picture’ understanding of the overall rhetorical purpose of the texts in the genre. Once the overall rhetorical purpose is understood, it is then necessary to look at the *function* of each text segment and evaluate what its local purpose is. This can be rather difficult. Move types



need to be distinctive. Multiple readings and reflections on the texts are needed before clear move types, with their defining function(s), emerge. During this process, one needs to look for any common functional and/or semantic themes represented by the various text segments that have been identified, especially those that are in relative proximity to each other or often occur in approximately the same location in various texts representing the genre. These functional-semantic themes can then be grouped together, reflecting the various steps of a broader move type, with each move having its own functional-semantic contribution to the overall rhetorical purpose of the text.

In summary, the first step of the BCU Approach (Develop Communicative/Functional Categories) is realized in a move analysis through this two-part process:

1a) Determine Rhetorical Purpose of the genre.

1b) Determine Rhetorical Function of each text segment in its local context:

- identify the possible move types of the genre;
- group functional and/or semantic themes that are either in relative proximity to each other or often occur in similar locations in representative texts. These reflect the specific *steps* that can be used to realize a broader move;
- conduct pilot-coding to test and fine-tune definitions of move purposes;
- develop coding protocol with clear definitions and examples of move types and steps.

TABLE 2. BCU Approach: *Application to corpus-based move analysis*

<i>Required step in the corpus-based discourse analysis (see Table 1)</i>	<i>Application to move analysis</i>
1. Communicative/functional categories	Determine rhetorical purpose of the genre. Determine rhetorical function of each text segment in its local context: <ul style="list-style-type: none"><li>• identify the possible move types of the genre;</li><li>• group functional and/or semantic themes that are either in relative proximity to each other or often occur in similar locations in representative texts. These reflect the specific <i>steps</i> that can be used to realize a broader <i>move</i>;</li><li>• conduct pilot-coding to test and fine-tune definitions of move purposes;</li><li>• develop coding protocol with clear definitions and examples of <i>move types</i> and <i>steps</i>.</li></ul>
2. Segmentation	Segment full set of texts into moves. (This step in a move analysis is done concurrently with Step 3: Classification.) <ul style="list-style-type: none"><li>• run inter-rater reliability check to confirm that there is clear understanding of move definitions and how <i>moves/steps</i> are realized in texts;</li><li>• identify any additional <i>steps</i> and/or <i>moves</i> that are revealed in the full analysis;</li><li>• revise coding protocol to resolve any discrepancies revealed by the inter-rater reliability check or by newly ‘discovered’ <i>moves/steps</i>, and re-code problematic areas.</li></ul>
3. Classification	Classify all moves by move type.
4. Linguistic analysis of each unit	Conduct linguistic analysis of all moves representing each move type.
5. Linguistic description of discourse categories	Describe move types in terms of the linguistic features of the moves.
6. Text structure	Analyze the move structure of each text in terms of move types.
7. Discourse organizational tendencies	Describe the corpus of texts in terms of typical and alternate move structures.

5.1.1 Application to the birthmother letters study

Determining the overall rhetorical purpose of the letters in the birthmother letter corpus was fairly easy: it is to convince an expectant mother, considering an adoption plan for her unborn child, to consider the couple as her child’s future parents. Identifying specific move types was more challenging because it required reading all the letters carefully and keeping track of what writers

were trying to accomplish rhetorically.

Having a trained assistant to discuss and debate move types and structures greatly facilitated this process. The hardest part was making the distinction between what was a unique move type and what was really a step (or strategy) of an already identified move type, but following the process for determining the rhetorical function of each segment discussed above, a preliminary move structure for the genre emerged, which was used and refined as described in the strategies below.

Table 3 provides an example of how one move type and its steps were defined, along with an example of how this move type was realized in letters. Move Type 3, Describing the couple's history before marriage, has the overall purpose of showing a birthmother what the lives of the couple writing the letter were like before they became a couple. Originally, this move type was simply defined as 'describing the couple's history', but after analyzing the whole corpus and reflecting more on the rhetorical function of this move type, it became clear that it should be narrowed to describing the individual members of the couple 'before marriage'. The three steps of this move type represent the three different ways this rhetorical purpose is typically realized in the letters.

The letter writers may describe their childhoods (Step 3A), describe their adult lives before they were married (Step 3B), or describe how they became a couple (Step 3C). Cohen (2007) provides a thorough discussion of how these move types were identified and refined during the analysis process.

TABLE 3. *Descriptions and examples of move type 3 and steps*

Move Type 3: *Describing the couple's history before marriage*

This move type serves the general function of showing a potential birthmother what the lives of each person were like before they became a couple. It is realized through one or more of the following three steps:

Step 3A: Describing their childhoods: In this step, the couples discuss what family life was like for them while growing up. The values they grew up with and the memories that they cherish are included in this step.

Example: I was lucky to grow up with all four of my grandparents nearby. Every summer we would head south for family vacation. I was especially close to my mom's mother. I spent many happy times at their lakeside condominium, sailing and riding in the speedboat on warm summer days. We went ice fishing on the lake in the winter, and once we went to the World's Fair in Tennessee. But it was the quieter times we spent together that I treasure most. My grandma taught me how to sit like a lady and how to hold eating utensils properly. She read me a story before bed each night. (S3)

Step 3B: Describing their adult lives before they were married: Descriptions of what their lives were like after high school, during college, and before they found each other comprise this step.

Example: I have a degree in fashion merchandising and used to be a store manager for a clothing store. I started college full time when I was 20. I worked hard to finish school early, I took my pre-med classes in college year round so that I could start graduate school After six years of hard work, I finally graduated with a doctor's degree. (S6)

Step 3C: How they became a couple: This step details how the couple met, started dating, and eventually married. It also encompasses details from their weddings.

Example: The first time I saw Steve, I was sitting with a group of friends when one of them pointed him out across the room. One of the first things I noticed was his smile and how his face lit up a room. He wasn't even aware of how his smile drew me to him. The day Steve proposed to me, he didn't ask me to marry him but asked, 'Will you please be my wife?' That day I felt like the luckiest person to have found such a man whom I love, admire and respect. (S1)

5.2 SEGMENTATION AND CLASSIFICATION

When one is ready to segment the texts in a corpus into moves and classify those into move types (steps 2 and 3 of the *BCU Approach*), it is best to begin first with a pilot coding, ideally with at least two coders. Because coders are seeking to under-stand the functional-semantic purposes of text segments, this segmentation and classification must be done by hand. Initial analyses are then discussed and fine-tuned until there is agreement on the functional and semantic purposes that are being realized by the text segments. This process is greatly facilitated by a coding protocol of move type and step features for the

genre, with clearly de-fined purposes and examples, which is developed first (see section 5.1 earlier and Appendix 1), although this protocol is inevitably refined during the segmentation and classification stages.

For a corpus-based move analysis, the coding protocol is then applied to the full set of texts so that each move in each text is classified by move type. Inter-rater reliability should be checked to confirm that there is agreement on what the move types are and how they are realized by text segments. (See p. 35 in Biber et al. 2007, for a fuller discussion on inter-rater reliability.) At this point, it may be necessary to resolve any discrepancies through further discussion and analysis, and then re-code problematic texts. It is also common that additional steps or even move types will be discovered during the analysis of the full set of texts.

Some move structures can prove fairly complex. For example, Bhatia (1998: 100) has noted that fundraising discourse ‘offers a large variety of creative options’. In other words, some genres, especially dynamic and persuasion-oriented ones like fundraising letters, may have obligatory, typical, and optional move elements, and move types may not necessarily occur in a fixed order (see e.g. Biber et al., 2007, chapter 3). Nevertheless, a move structure for a genre can still be identified by working through the general process outlined above.

In summary, the second (Segment each text into discourse unit) and third (Classify functional type of each discourse unit in each text of the corpus) steps of the *BCU Approach* typically are done concurrently in a move analysis and include the following:

2) Segment full set of texts into moves:

- run inter-rater reliability check to confirm that there is clear understanding of move definitions and how they are realized in texts;
- identify any additional *steps* and/or *moves* that are revealed in the full analysis;
- revise coding protocol to resolve any discrepancies revealed by the inter-rater reliability check or by newly ‘discovered’ *moves/steps*, and re-code problematic areas.

3) Classify each move by move type.

### 5.2.1 Example from the birthmother letters study

While segmenting each of the birthmother letters into discourse units was relatively straightforward – that is, it usually was not that hard to determine where one move stopped and another started – it was not always such a simple task to confidently identify the move type the segment represented. In order to identify the move type, the communicative purpose of each text segment had to be considered. Some move types are fairly obvious; for example, all letters contained a section that provided contact information for the potential birthmother or her attorney, and this move was easy to label as Move Type 9 (*Concluding letter with contact information*). The move type of other text segments, however, needed to be determined by the context of the letters. In the following example, two text segments that discuss a couple's pets appear in the same letter, but they were identified as representing two different move types based on the communicative purpose.

Excerpt 1: I, Kate, was working one day when a client came into the office and was telling me about a puppy that had been abandoned on the side of a busy highway. I left to find this puppy and located him along the road. He was shaking so hard when I picked him up in my arms. He was infested with fleas and very sick, so I took him to the vet where he was treated and released. (S1)

Within the context of the letter, this segment's purpose was not so much on the animals themselves, but is used to show the generous, warm character of the letter writer; consequently, it was labeled as Move Type 6 (Profile of the couple – Step 6B: Character and values).

Excerpt 2: Six years ago, we decided to adopt a dog from the Humane Society. This is how Abby became a part of our family. The first year she chewed everything in sight with her favorite being the remote control (another reason we don't mind the deer eating our flowers!).

Two years later Cosmo joined our family. He has been a part of our family ever since then. They are both a big part of our lives and are very loving and playful around others. (S1)

This excerpt from the same letter focused on describing the pets in the family and so was identified as Move Type 7 (Physical environment – Step 7B: Pets).

While both of these moves had animals as their topic, it was important to look closely at what the purpose of the text segments, within their contexts, were. In this example, the focus of the first text segment was on the action of the

couple, and the focus of the other was on the pets themselves.

During the process of segmenting all the texts into moves and identifying move types, and as a result of conversations initiated by discrepancies in the inter-rater reliability checks, the initial protocol of move types evolved, with some of the move types combined or reconfigured and two steps re-evaluated as being distinct move types because of their importance to the communicative purposes of the letters. The final protocol of move types is provided in Appendix 1. A fuller description of how the move types were refined for the birthmother letters is provided in Cohen (2007).

### **5.3 LINGUISTIC ANALYSIS OF EACH UNIT AND LINGUISTIC DESCRIPTION OF DISCOURSE CATEGORIES**

The goal of the *BCU Approach* for the corpus-based analysis of discourse organization is to go beyond simply segmenting texts into well-defined discourse units (in this case, moves); the desire is also to analyze the linguistic characteristics of each individual discourse unit and each discourse unit type (i.e. the move types), to determine distinctive linguistic characteristics of the units and their distribution. Although they are defined in functional terms, move types are realized through linguistic features, including word choice, phrase types, and grammatical features (e.g. tense, aspect, voice).

For example, in Biber et al. (2007), a corpus of direct-mail fundraising letters was first analyzed for the rhetorical move structure common to the texts in this genre. Using a grammar tagging program, all of the words in each move were then linguistically identified (tagged); once this was done, all of the moves in each of the move types were then analyzed for their use of grammatical stance devices (see Biber et al., 1999), in particular: 1) stance adverbials, 2) stance complement clauses (specifically ‘that’ and ‘to’ clauses), 3) modals, 4) premodifying stance adverbs (e.g. ‘I’m *so* happy for you’), and 5) stance nouns followed by prepositional phrases.

Because non-profit direct mail letters are overtly persuasive in nature, there is little question that stance plays an important role in this genre; consequently, exploring stance structures seems a reasonable investigation in the effort to better understand the linguistic structures used in direct-mail letters. In fact, Biber et al. (2007) show that by looking at the use of these stance devices across the move types within this single genre, it is possible to describe each different move type by the combination of one or more of these grammatical

stance devices, with no two move types using the same set of stance structures.

In summary, the fourth (Linguistic analysis of each unit) and fifth (Linguistic description of discourse categories) steps of the *BCU Approach* typically are reflected in a move analysis through the following two strategies:

- D. Conduct linguistic analysis of move features.
- E. Describe move types in terms of the linguistic features of the moves.

### 5.3.1 Example from the birthmother letter study

Two hypotheses we had about language use in the birthmother letters were, first, that pronoun usage would likely differ across the different move types, and, second, that writers of successful birthmother letters would use pronouns differently than the writers of the unsuccessful letters. One structure we looked at specifically was the combination of the possessive pronoun ‘our’ used with the nouns ‘baby’ and ‘child’. We used *Wordsmith* (Scott, 2004) to first identify all uses of these pronoun structures in all moves across all move types. We then looked at how the overall use of these structures varied across move types. Move Type 6 (Profile of the couple) had by far the most frequent occurrence of these structures, in particular Step 6B (Character and values), which is when the couple talks about their character and values as they describe more broadly who they are as a couple. It turns out that for both the successful and unsuccessful letters, the couples used the structure ‘our baby’ and/or ‘our child’ in Move Type 6 (Profile of the couple) in a similar way: to describe how they envision being a parent and what they want for a child. The following examples reflect this purpose:

S2: We are eager to include our baby in all of the simple things . . .

S6: We want what is best for our child, including a good education . . .

U1: . . . that Becky will be a great mother to our child. She has such a creative mind . . .

U18: . . . with each other as a couple. When our baby arrives, we look forward to . . .

However, there was a very notable difference in the rate that the successful and unsuccessful writers used these structures. Writers of the successful letters used ‘our baby’ or ‘our child’ at a rate of 5.2 times per letter in Move Type 6 (Profile of the couple), while the writers of the unsuccessful letters used those



two structures at a rate of 3.0 times per letter in Move Type 6, more than 40 per-cent less frequently. The result is that within the unsuccessful letters, there are fewer occurrences that involve an explicit discussion about the vision a couple has in parenting their adoptive child. Overall, this possessive personal pronoun plus noun structure functions as a way of showing a reader what the adoptive child's life will be like in a family. By more frequently using 'our child' and 'our baby' as they talk about what their life is and will be like, the letter writers help the expectant mother more easily envision her child in a particular environment, and she can more easily see a couple's intentions. Obviously, frequent use of this structure is not the only basis on which a birthmother decides which couple she wants to have adopt her child, but it does reflect how language is used differently in the successful and unsuccessful letters.

While the above example only briefly describes one of the linguistic features (first person plural pronoun usage) of a single move type (Move Type 6) in this genre, using the *BCU Approach* similar analyses of all move types were conducted to provide a more thorough description of the distinct linguistic characteristics of each move type. Further description of the linguistic features of the different move types in birthmother letters is given in Cohen (2007).

#### **5.4 TEXT STRUCTURE AND DISCOURSE ORGANIZATIONAL TENDENCIES**

When a move analysis is completed, the move structure should represent the 'rhetorical movement' (Swales, 1990: 140) of the functional-semantic purposes of the text segments that make up the genre. There are two levels of analysis that can be done here. First, the discourse organization of individual texts in the corpus can be analyzed, which permits us to describe the internal discourse organization (i.e. move structure) in terms of specific move types. The second level of analysis is the investigation of what the 'preferred' move structure is for the corpus as a whole; that is, what are the typical distinctive patterns of discourse that occur when looking at the texts in the corpus as a group, and how much variation is there? An excerpt from Biber et al. (2007: 54–5) describing a move analysis on direct mail letters, and the comments that follow, will help illustrate the process:

Using the rubrics given . . . outlining the rhetorical moves of the direct mail letters. . . two raters hand-coded the rhetorical moves . . . in all 242 letters in the corpus. As noted [earlier], individual moves often reappeared throughout a letter, and each appearance was counted as a distinct occurrence; as a result a single move type could occur multiple times. Inter-rater reliability was calculated at 84%, with all

discrepancies reconciled through discussion. The vast majority of discrepancies that occurred between the two raters resulted from initial disagreement as to where one move ended and the next started, not as to the presence of a particular move. This inter-rater reliability is quite good, since, as Bhatia notes, there are sometimes ‘cases which will pose problems and escape identification or clear discrimination, how-ever fine a net one may use. After all, we are dealing with the rationale underlying linguistic behavior rather than its surface form’ (Bhatia, 1993a: 93). Once all of the moves were agreed upon and marked, each letter was then tagged to indicate the start and stop of each move in each text.

The sequence of each move type . . . for each text was also noted. This allowed for the tracking of the total frequency of each move type in the corpus, their relative locations in each letter (e.g. first, second, third), what other move types a move most commonly occurred with, how frequently a move was embedded in another move, and how frequently a move type occurred in the body of the text as opposed to in a P.S.

Once all of the moves were tagged by type and relative location, Biber et al. (2007) were able to make specific observations about how moves are used within the genre. In fact, one result of this study was a set of three prototypes reflecting common move structures for direct mail letters, based on how frequently different moves tended to occur in the corpus as well as their typical position in the letter in relationship to each other and their mean length in terms of words.

In summary, the sixth (Text structure) and seventh (Discourse organizational tendencies) steps of the *BCU Approach* typically are reflected in a move analysis with these two strategies:

- F. Analyze the move structure of each text in terms of move types.
- G. Describe the corpus of texts in terms of typical and alternate move structures.

### **5.4.1 Example from the birthmother letter study**

Appendix 2 provides excerpts from a birthmother letter analyzed by move type. Once the move structure for each of the letters in the corpus was identified, it was then possible to look at what the typical move structure patterns are for these types of letters. Table 4 shows which move types tended to be more predictable, occurring in all or most of the letters in the corpus, and which move types appeared to be more optional, occurring in only some of the letters. By keeping track of where move types occur relative to other move types in a letter, we were also able to make observations about where in birthmother letters move types occurred. For example, Table 4 shows that for

the successful letters, seven of the 10 move types (1, 2, 3, 6, 7, 8, 9) occurred at least once in 90–100 percent of the letters, indicating that these are expected move types in birthmother letters. On the other hand, three of the move types (4, 5, 10) are clearly optional, occurring no more frequently than in 20 percent of the successful letters.

TABLE 4. *Appearance of and average total number of words in Move Types in successful & unsuccessful birthmother letters*

Letter Type/ Move Type	1	2	3	4	5	6	7	8	9	10
Successful letters (n = 20)	20/20 100%	20/20 100%	19/20 95%	16/20 80%	3/20 15%	20/20 100%	18/20 90%	18/20 90%	20/20 100%	4/20 20%
Avg. total # of words/letter	21	192	319	96	80	1,053	140	184	58	7
Unsuccessful letters (n = 26)	24/26 93%	26/26 100%	25/26 97%	22/26 85%	16/26 62%	26/26 100%	22/26 85%	23/26 89%	25/26 100%	1/26 4%
Avg. total # of words/letter	15	158	166	146	21	955	89	142	64	6

Average total number of words per move type for successful and unsuccessful letters also varies notably. (Letters may have multiple occurrences of a particular move type; word count totals represent the total of all occurrences of a move type.) For example, while both types of letters use Move Type 3 (Describing the couple’s history before marriage) at a comparable rate – 95 percent and 97 percent respectively – the average total number of words used in that move type in successful letters is 92 percent greater (319 words per letter when used) than for unsuccessful letters (166 words per letter when used). This difference in average number of words for Move Type 3 suggests that one of the advantages of the successful letters is that they give the birthmother a better sense of who the prospective parents are as individuals than the unsuccessful letters typically do.

The biggest discrepancy between move usage rate for successful and unsuccessful letters is with Move 5 (*Role of Religion*). While only 15 percent of the successful letters made explicit reference to the role of religion in the couple’s lives, 62 percent of the couples who wrote unsuccessful letters made explicit reference to the role of religion in their lives.

Closer analysis of where the move types occurred in the letters showed that while Move Type 1 (Introduction) always occurred first in the letters, Move Type 2 (Thanking, empathizing and/or reassuring expectant mother) could

appear in multiple places, including right after Move Type 1, at the end of the letter as one of the last moves, and even in the middle of the letter surrounded by a variety of move types. Similarly, Move Type 6 (Profile of the couple) typically occurs early in the letter and is most commonly followed by Move Type 7 (Physical environment). This information can be used to provide different examples of ‘prototypical’ letters that reflect typical organization structures for successful birthmother letters.

Again, the above discussion is intended to be an example of the type of analysis and description that is done as part of the *BCU Approach* to discourse analysis. A more thorough analysis and description of the move structure the birth-mother letter genre is provided in Cohen (2007).

## **6. Final thoughts on using a corpus-based approach to move analysis**

Discourse analysis in general, and move analysis in particular, has typically been a qualitative endeavor, with studies focusing on only a few texts. In contrast, a corpus-based approach requires analysis of a well-designed ‘representative’ collection of texts of a particular genre. These texts are encoded electronically, allowing for more complex and generalizable research findings, revealing linguistic patterns and frequency information that would otherwise be too labor intensive to uncover by hand (Baker, 2006). That is not to say that a corpus-based approach is simply a quantitative approach. Corpus-based discourse analysis depends on both quantitative and qualitative techniques. Even with a corpus-based approach, the moves and move types in each text must first be identified and tagged individually by the researchers making qualitative judgments about the communicative purposes of the different parts of a text; and even once quantitative data are run, the results must still be interpreted functionally. As has been noted previously, ‘Association patterns represent quantitative relations, measuring the extent to which features and variants are associated with contextual factors. However, functional [qualitative] interpretation is also an essential step in any corpus-based analysis’ (Biber et al., 1998: 4).

To summarize, a corpus-based approach to move analysis differs from the ‘traditional’ approach in the following ways:

a) Analyses are done on a relatively large representative collection of texts

from a particular genre. In the birthmother letters study, we had a corpus of 46 letters. While a corpus this size is sufficient to get meaningful data, the larger the corpus the easier it is to see and validate linguistic patterns and trends (Baker, 2006).

b) All texts are electronically encoded to allow for computerized counts and calculations using different programs and software packages. In our birthmother letter study, because the data were electronically encoded, we were able to use Wordsmith (Scott, 2004) to analyze the data. Other concordancers, such as Monoconc Pro or Concordance, could also be used.

c) Analysis of the linguistic characteristics of specific move types can be easily done in order to provide details about how different communicative purposes are realized linguistically. While one could do a move analysis of a single text (see Mann and Thompson, 1992), it only becomes possible to describe the typical linguistic characteristics of move types through a corpus-based approach. Move types, and their component steps, are identified by the functional and semantic purposes that they have. Consequently, because different moves have different functional and semantic purposes, it seems reasonable to expect that move purposes will be realized through variations in linguistic features. With computers, much more interesting and comprehensive linguistic analyses can be undertaken than can be done by hand. In fact, co-occurrence of linguistic features and how features interact with each other in a move to perform a particular communicative purpose can be easily studied. It is often more informative and useful to study the distribution and co-occurrence of many features of language at once, rather than considering the distribution and function of individual features singly. Computer driven, corpus-based approaches allow us to do this. It was only because we had a reasonably large corpus of birthmother letters in our study that had been coded and tagged by move type were we able to take full advantage of the sorting features of Wordsmith (Scott, 2004) to identify and analyze the use of linguistic structures (like pronouns) within particular move types. The ability to analyze the linguistic realizations of move types is an important step forward in understanding how language is used for specific communicative purposes.

d) In addition to conducting the traditional move analysis, quantitative counts permit the discussion of general trends, including the relative frequency of particular move types. Once moves in a corpus have been coded, a variety of descriptive counts can be made. The most obvious of these are the overall frequency of occurrence of each move type in the corpus, and the average

length in words of each move type. Statistics like those presented in Table 4 for the birthmother letter study allow us to make a clear determination as to whether a particular move type is obligatory, expected, or merely optional.

e) Prototypical and alternate patterns of move type usage can be identified and mapped. A computer can be used to count not only the presence of each move type for each text but also to keep track of their positions relative to each other (e.g. first, second, third), what other move types each most commonly co-occur with, how frequently a move is embedded in another move. With statistics on move frequencies and lengths, as well as descriptions of where in the genre a move type tends to occur and how one move type typically relates to another, a key advantage of a corpus-based approach can be realized: the ability to develop genre prototypes. Prototypes are particularly valuable in educational and training contexts to help novices learn to understand and produce a genre that is new to them. Cohen (2007) provides a prototype of a typical birthmother letter that can be used to assist couples as they write their own letters.

To summarize, this article introduced a general seven-step corpus-based approach to discourse analysis, then showed how it can be applied to move analyses to provide a more thorough description of typical patterns of discourse organization, including a better description of the linguistic features of move types. A corpus-based approach – the *BCU Approach* – for conducting a move analysis is described, which significantly enhances the value of this often used (and misused) methodology, while at the same time providing badly needed guidelines for a methodology that lacks them.

#### ACKNOWLEDGMENT

This article draws on material originally published in *Discourse on the Move: Using Corpus Analysis to Describe Discourse Structure*, co-authored by Douglas Biber, Ulla Connor, and Thomas Upton, published in 2007 by John Benjamins, Amsterdam, and is used with permission of the publisher and all co-authors.

#### APPENDIX

##### APPENDIX 1. Final coding protocol of move types in birthmother letters

###### Move 1: Introductions

- Move 2: Thanking, empathizing and/or reassuring expectant mother
- Move 3: Describing the couple's history before marriage
  - Step 3A: Childhoods
  - Step 3B: Adult lives
  - Step 3C: How they became a couple
- Move 4: Why couples want to adopt
  - Step 4A: Fertility reasons
  - Step 4B: Connections to adoption and respect for birthmother
- Move 5: Role of religion
- Move 6: Profile of the couple
  - Step 6A: Hobbies and travels
  - Step 6B: Character and values
  - Step 6C: Economic situation
- Move 7: Physical environment
  - Step 7A: Description of the home and surroundings
  - Step 7B: Pets
- Move 8: Support systems
  - Step 8A: Other children in the household
  - Step 8B: Family and friends
- Move 9: Concluding letter with contact information
- Move 10: Describing the letter's structure

## APPENDIX 2 .Excerpts from a Successful Letter (S2) Tagged for Moves

<1> Dear Expectant Mother:

We are Joe and Barbara. </1> <2> It is hard to put into words what we really want to express to you. We want you to know who we really are, what we think, and how we feel from our hearts</2>

<10> Why the Italics and Bold.

We wanted you to hear both of our voices in this letter. We thought it would be fun to use Italics when I (Barbara) am speaking, and regular type when it's me (Joe). There will be times when we both want to say the same thing, so when it appears in bold black type, we're both talking to you. </10>

<3C > About us:

My father and Joe worked together, and through the wonderful interference of my mom, we met at the company Christmas party 13 years ago. Barbara and I quickly became inseparable and we were married 4 years later. </3C> <6B>We are now in our early 30s and are excited to become parents through adoption.</6B>

<6C> Like most families today, Joe and I both work. I'm lucky to have a job that I enjoy, and a boss that understands how important it is to spend time with your children. I work as a dental assistant with advanced skills and responsibilities. I am home most days between noon and 2:00 pm, so that leaves a lot of time to spend with baby. I still work for the same Construction Company as a union bricklayer. I have good benefits, a nice income and I enjoy what I do. I take pride in my work. </6C>

<6A> Hobbies and Interests:

Joe and I enjoy the outdoors especially in the summer. While Barbara gets her hands dirty in her flower gardens, I take care of the lawn. I also enjoy the challenge of target shooting. It takes patience, precision, and skill to produce accurate results. I not only

enjoy gardening, but also, sewing, crafts, or curling up with a good book. One of my favorite pastimes is going to auctions and antiquing. </6A> <3A>As a child, I can still remember standing in a farmyard or old barn between my mom and grandma, and watching the bidding war begin. </3A>

<6B> As we stated earlier, summer is our favorite season. There is so much to do and see. </6B> <6A> We take great pleasure in sitting on the porch swing, on warm summer nights, and listening to the frogs and crickets sing. On hot days we take Jack (our dog) to my grandparents pool to cool off and visit with the family. There are always kids (big and little) splashing around and having fun. </6A> . . .

<8B> The Family:

We have a very extensive family consisting of grandparents, parents, uncles, aunts, and cousins. The whole family welcomes the opportunity of giving their love and support to a new child. </8B> <7B> Up until now, one of our most important family members is white, fluffy, and four legged. She's our silly, little dog, Jack. </7B> . . .

<6B > Personal Thoughts:

I'm one of the luckiest ones to have found my soul mate on the first try. I knew I loved Joe after our first date, and I wanted to marry him on our second. Joe is the kind of person you strive to become close to. He has an honest, forthcoming feel about him . . . When you are asked the question, "If you could turn back time would you do things differently?" I can honestly say I wouldn't. I have a great life and a wonderful relationship with my wife. I know that sounds storybook, but it's true. Barbara makes my life what it is today. Without her I wouldn't be as happy, and fulfilled. Sharing my life with Barbara has given me a lot to be thankful for. I can't wait to be able to see her, share that kind of love and influence with a child. My philosophy in life is, "Being rich doesn't always mean having a lot of money." </6B> . . .

<2> We want to thank you for taking the time to learn a little bit about us. </2>

<9> We would very much like to talk to you in person. You can reach us through our toll free number at 1-234-567-89810. We don't want to miss your call, so please contact us after 4:00 p.m . . . </9> <2> Please know that if you call, no one will try to pressure or influence you in any way. We want to provide you with as much information as possible. </2>

<9> We wish you and your child the best, </9>  
[Signatures]

## REFERENCES

- Baker, P. (2006) *Using Corpora in Discourse Analysis*. London: Continuum.
- Bhatia, V. (1993a) *Analyzing Genre: Language Use in Professional Settings*. London: Longman.
- Bhatia, V. (1993b) 'Simplification vs. Easification: The Case of Legal Texts', *Applied Linguistics*4(1): 42-54.
- Bhatia, V. (1998) 'Generic Patterns in Fundraising Discourse', *New Directions for Philanthropic Fundraising* 22: 95-110.
- Biber, D., Connor, U. and Upton, T. (2007) *Discourse on the Move: Using Corpus Analysis to Describe Discourse Structure*. Amsterdam: John Benjamins.
- Biber, D., Conrad, S. and Reppen, R. (1998) *Corpus Linguistics Investigating Language Structure and Use*. Cambridge: Cambridge University Press.
- Biber, D., Johansson, S., Leech, G., Conrad, S. and Finegan, E. (1999) *Longman Grammar of Spoken and Written English*. London: Pearson Education.
- Cohen, M.A. (2007) 'Dear Birthmother: A Linguistic Analysis of Letters Written to Expectant



- Mothers Considering Adoption', unpublished master's thesis, Indiana University, Indianapolis, available online at: [<http://hdl.handle.net/1805/1114>].
- Connor, U. (1996) *Contrastive Rhetoric. Cross-Cultural Aspects of Second-Language Writing*. Cambridge: Cambridge University Press.
- Connor, U. and Upton, T. (2004) 'The Genre of Grant Proposals: A Corpus Linguistic Analysis', in U. Connor and T. Upton (eds) *Discourse in the Professions: Perspectives from Corpus Linguistics*, pp. 235–56. Amsterdam/Philadelphia, PA: John Benjamins.
- Hoey, M. (1986) 'Overlapping Patterns of Discourse Organization and Their Implications for Clause Relational Analysis in Problem-Solution Text', in C.R. Cooper and S. Greenbaum (eds) *Studying Writing: Linguistic Approaches*, pp. 187–214. Newbury Park, CA: Sage.
- Kanoksilapatham, B. (2005) 'Rhetorical Structure of Biochemistry Research Articles', *English for Specific Purposes* 24: 269–92.
- Mandler, J.M. and Johnson, N.S. (1977) 'Remembrance of Things Parsed: Story Structure and Recall', *Cognitive Psychology* 9: 111–51.
- Mann, W. and Thompson, S. (eds) (1992) *Discourse Description: Diverse Linguistic Analyses of a Fund-Raising Text*. Amsterdam/Philadelphia, PA: John Benjamins.
- Nwogu, K. (1991) 'Structure of Science Popularizations: A Genre-Analysis Approach to the Schema of Popularized Medical Texts', *English for Specific Purposes* 10(2): 111–23.
- Posteguillo, S. (1999) 'The Schematic Structure of Computer Science Research Articles', *English for Specific Purposes* 18(2): 139–58.
- Samraj, B. (2002) 'Introductions in Research Articles: Variation across Disciplines', *English for Specific Purposes* 21: 1–17.
- Schiffrin, D., Tannen, D. and Hamilton, H. (eds) (2001) *The Handbook of Discourse Analysis*. Oxford: Blackwell.
- Scott, M. (2004) *Oxford Wordsmith Tools Version 4*. Oxford: Oxford University Press.
- Swales, J. (1981) *Aspects of Article Introductions*. Birmingham: University of Aston.
- Swales, J. (1990) *Genre Analysis: English for Academic and Research Settings*. Cambridge: Cambridge University Press.
- Thompson, S. (1994) 'Frameworks and Context: A Genre Based Approach to Analyzing Lecture Introductions', *English for Specific Purposes* 13(2): 171–86.
- Upton, T. (2002) 'Understanding Direct Mail Letters as a Genre', *International Journal of Corpus Linguistics* 7(1): 65–85.
- Upton, T. and Connor, U. (2001) 'Using Computerized Corpus Analysis to Investigate the Textlinguistic Discourse Moves of a Genre', *English for Specific Purposes: An International Journal* 20: 313–29.
- van Dijk, T. (1988) *News as Discourse*. Hillsdale, NJ: Erlbaum.
- Williams, R. (1999) 'Results Section of Medical Research Articles: An Analysis of Rhetorical Categories for Pedagogical Purpose', *English for Specific Purposes* 18(4): 347–66.