A Cognitive Approach to Judicial Opinion Structure: Applying Domain Expertise to Component Analysis

Jack G. Conrad
Research & Development
Thomson Legal & Regulatory
St.Paul, MN 55123 USA
Jack.Conrad@WestGroup.com

Daniel P. Dabney
West Online Research
West Group
St.Paul, MN 55123 USA
Daniel.Dabney@WestGroup.com

Abstract

Empirical research on basic *components* of American judicial opinions has only scratched the surface. Lack of a coordinated pool of legal experts or adequate computational resources are but two reasons responsible for this deficiency. We have undertaken a study to uncover fundamental components of judicial opinions found in American case law. The study was aided by a team of twelve expert attorney-editors with a combined total of 135 years of legal editing experience. The scientific hypothesis underlying the experiment was that after years of working closely with thousands of judicial opinions, expert attorneys would develop a refined and internalized schema of the content and structure of legal cases. In this study participants were permitted to describe both concept-related and format-related components. The resultant components, representing a combination of these two broad categories, are reported on in this paper. Additional experiments are currently under way which further validate and refine this set of components and apply them to new search paradigms.

Keywords

document structure, cognitive models, case law analysis

1. INTRODUCTION

This experiment grew out of our need to establish a knowledge-base for related legal corpus research, including studies in textual abridgement. Of West Group's thousands of full-text databases representing terabytes of information, the majority focus on the legal domain. To assist our efforts in developing new or enhanced search tools for the legal or business professional, it was essential to have a set of reliable, empirically-corroborated notions about the structure and relationships in case law documents. As Deedman and Smith have underscored:

> It is important to realize that no logical overall structure of this area of law had previously existed. Most nonlawyers find the lack of organization difficult to believe and somewhat disconcerting. However, this formlessness is characteristic of the Anglo-American common (case) law tradition, as opposed to the European civil law which is codified. Basically, a body of case law consists of a large collection of cases that have been accumulated over time, often hundreds of years. The cases are supplemented by commentaries found in the legal literature. An area of common law grows by process of accretion, somewhat like a coral reef, as new cases are decided and added to its corpus [11].

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
ICAIL-2001 St. Louis, Missouri USA
Copyright 2001 ACM 1-58113-368-5/01/0005 ...$5.00.

If there exists an apparent lack of structure at the corpus level, such absence of universal form is even more pronounced at the individual case level. As many professors of law will assert, no one style of judicial opinion writing fits all judges.[1] Moreover, the surface-level structure that one finds in a body of cases is nearly as varied as the subject matter it addresses. This is despite the fact that over time judges and others have tried to establish effective ways of writing opinions [1].

With a sizable staff of highly-skilled attorneys employed to identify and summarize points of law, it seemed reasonable that over time, these professionals would form some stable notions of what components comprise a judicial opinion. Thus began our efforts to tap into this schemata of common law [22].

Our experimental objectives included learning more about the discourse-level or 'macrostructure' textual features of judicial opinion documents [27]. Another central motivation behind this research was to determine to what extent editors would achieve agreement and consistency in their analysis, since if human experts could not agree, then little success could be expected when automatically handling these case law components. For large publishing enterprises or online service vendors which may invest hundreds of thousands of person-hours each year in the review and editing of judicial opinions, such an endeavor could result in editorial savings as well as new means of accessing judicial opinions. Longer range goals for such research include automating the process of identifying and connecting such added features. The results presented in this paper are suggestive and promising, though not definitive. However extensive our editorial resources, for practical reasons they were not unlimited.

Section 2 of this paper reviews related work in the area of document structure analysis. Our experimental methodol-

[1] In this paper, *judicial opinion* is used to denote the statement by a judge or court of the decision reached regarding a cause tried or argued before them, which expounds the law as applied to the case, and details the reasons upon which the judgment is based. By contrast, *case law* or *legal cases* refers to the collection of reported cases which forms a body of jurisprudence (i.e., the law of a particular subject formed by the decided cases) and is distinct from statutes and other sources of law.

ogy is described in Section 3. Section 4 presents the results of our study, while Section 5 examines the conclusions we are able to draw from these findings. Lastly, Section 6 presents an overview of the new directions this research is taking.

2. RELATED WORK

A great deal has been written about the lexical and grammatical aspects of legal discourse [9, 20] and models of argumentative discourse have been widely studied [3, 7]. Yet a number of these models have been developed without extensively consulting the human practitioners of their domains. In other words, they appear to be built from the machine down rather than from the human expert up.

Paice treats the natural language understanding challenge faced by text abstractors as essentially an information extraction task which is performed during automatic indexing [21]. In this work, an assortment of candidate "fillers" are considered for various slots in a semantic frame. Concept names are then selected for these slots by examining the candidates and associated weights. Others have imposed synthetic structures on textual documents in other ways, for instance, by segmenting a text into multi-paragraph units—thereby being prepared to respond to a new regiment of queries capable of focusing on the subtopic domain [13, 12].

By contrast, a number of researchers have performed extensive analytical studies on the existing internal structure of various modes of natural language discourse. These have included general natural language prose [27, 14], the 'grammar' of discourse [17], as well as exposition [6], argumentation in editorial discourse [3], news as discourse [29], and historical discourse and associated abstracts [25]. Most of these projects share the underlying philosophy that for too long fields such as linguistics have confined themselves to words, phrases, or sentences in isolation and have ignored the context in which they occur. Many of these studies owe their conception to the relatively new field of Discourse Analysis.

Discourse Analysis is a multidisciplinary approach to textual study and stems from a number of fields in the humanities and the social sciences—linguistics, anthropology, sociology, psychology, cognitive science, and others. Among the most important research in this area is that of van Dijk, whose seminal works on macrostructures proposed that one could examine the discourse of a domain in terms of its deeper sub-textual structure [27]. VanDijk, like Danet [9], tends to underscore the social and cultural dimensions of language use as much as the cognitive. Despite new developments in formal grammars, logic, and artificial intelligence, the field—methodologically and theoretically—still awaits broader application [28].

Liddy used some of these same propositions to explore the discourse-level structure of empirical abstracts and the existence of lexical clues which reveal such structure [16]. In a validation procedure, Liddy showed that professional abstractors do possess an internalized structure of empirical abstracts, whose components and relations were corroborated over the course of four exercises. The study was motivated by the supposition that many natural language features go unexploited in text-based information processing systems. Her experiments served as a model for some of the initiatives we report on here. However, focusing on abstract-like summaries of judicial opinions or the points-of-law contained in these cases, though more manageable, was not found to be as useful as dissecting the core opinions themselves.

A perhaps more computationally adaptable approach to textual study was introduced by Mann and Thompson in the form of Rhetorical Structure Theory (RST) [18]. RST has become one of the most widely acknowledged discourse theories in the last ten years. It offers an explanation of the coherence of texts, providing an analysis of the clauses and the relations between them due to their communicative content. Because it specifies the role of the observer of the text, it claims to provide a basis for examining the objectivity and subjectivity of the analysis. Central to the theory is the idea of rhetorical relation, which is the relation that binds two non-overlapping text spans called *nucleus* and *satellite*. The distinction between the two derives from the empirical observation that (a) the *nucleus* expresses more of the essence of the author's purpose, (b) the *satellite* supports it, and (c) the *nucleus* of a rhetorical relation is understandable independent of the *satellite*, but not the converse. A few exceptions to this rule exist, such as a *constraint*, a multinuclear case. RST thus provides a functional basis for studying the discourse-relevant forms or "discourse-markers" in texts, along with other formal correlates of discourse structure. RST has served as an effective tool that has been applied to a variety of contemporary computational linguistic problems, including summarization [19]. Yet because of the requisite training and inherent subjectivity—in terms of component identification and granularity—we opted to tap into the internalized schema of our own domain experts, without such externally imposed rhetorical relations.

Such a "Deep Structure" theory of law has been promoted by J.C. Smith [8]. His theory asserts that although legal decision making is basically a rule-governed activity, the rules that must be applied are beyond the "surface" rules of legal discourse.

More recently, Brüninghaus and Ashley have addressed the issue of indexing the knowledge represented in legal cases. They examined the utility of focusing on finer-grained textual units such as sentences [5]. They used an ID3 learning algorithm to determine what CATO-like factors may be associated with a given sentence [2]. Hence, a type is determined from a set of 26 abstract fact patterns. The usefulness of such building blocks outside of instructional environments, however, is an open question.

A number of limited-application legal expert systems have been designed and developed [24, 11, 12, 10]. Some claim to exploit the structural features of legal documents, yet they tend to circumscribe their domain such that the generality of their structured building-blocks remains unclear.

3. EXPERIMENTAL METHODOLOGY

In order to focus on the internal components of case law documents, we developed a three-phase study. With the help of an outside consultant, procedures were designed to describe and formalize the extensive knowledge of a team of legal-editing experts. All editors are attorneys and have considerable experience with judicial opinions. The average amount of legal-editorial experience for each editor is over 11 years (the median number of years being 13). Gender was equally represented on the team.

The phases consisted of recording the experts' notions about the common elements in a representative opinion, analyzing how reliably these experts could independently

identify and agree on these common elements, and, lastly, re-evaluating the resultant set of elements based on our findings as well as other factors such as the cost, utility, and coherence of the elements.

3.1 Phase I—Component Identification

In this phase, we conducted individual interviews of the participants in open-ended sessions lasting anywhere from 20 to 60 minutes. We asked each interviewee to list everything he or she could about "the common elements of judicial opinions in American case law." No other information was provided except to briefly answer clarifying questions (e.g., to focus on the raw opinion minus any added editorial enhancements such as summaries or notes about points of law). Each of the interviews resulted in a list of case components, annotated or qualified as the editor saw fit. Following the interviews a 90 minute focus group was held for the entire body of editors. During this time, the participants were asked the same basic question posed during the interviews. All responses were recorded on a white board in the front of the conference room. The session was also taped and a transcript was made for future reference. The participants themselves discussed whether a given element was a synonym for another element or whether it merited its own entry, and, when warranted, to what extent two elements might overlap. Near the end of the session, each of the participants was presented with a copy of his or her contributions produced during the earlier interview session and asked to ensure that each of his or her components was properly represented in the current resultant set [15].

In a separate *Q-sort* exercise, each of the participants was asked to sort a set of 52 cards, one card for each of the resultant components, into four piles [23, 26]. Each pile signified a given component's degree of "typicality."[2] We defined typicality as how regularly the component is present in an opinion. The exercise was repeated for "importance." We defined importance as how significant the component is (or could generally be expected to be) to a legal researcher. Later in a *Delphi-round* exercise, members were given the opportunity to modify any of their initial typicality or importance scores if they saw fit, after seeing the group's average score for each of the components [30, 4]. For this task, the subjects were presented with only their initial scores for the two categories along with the group's mean scores. The rationale for relying on this multi-step rating process was three fold. First, it ensured that each of the participants contributed equally to the final scores. Second, it gave each of the editors the chance to verify that they were viewing the four choice scale in a manner generally similar to their colleagues. Lastly, it gave each of the participants the opportunity to reassess their component scores when they deviated significantly from the group's mean values.

In the last exercise of this phase, the editors were asked to complete a questionnaire on the relationships between the components identified in the focus group.

[2] A four choice Likert scale [26] was selected by the focus group participants (using 1 for *most* typical or important and 4 for *least* typical or important) in order to give the respondents a reasonable number of choices and to produce a reliable set of data. By selecting an even number of choices, the middle-of-the-road 'no opinion' option was avoided. The scale was reviewed for appropriateness for each of the components being assessed.

3.2 Phase II—Trial Tagging of Components[3]

In preparation for the second phase, a number of the participants contributed to the construction of a glossary of the resultant components. (See Appendix) The editors were then given individual "tagging" surveys. Using a scale from 1 to 4, the participants were asked to assess how difficult it would be to (a) identify and (b) delineate boundaries for the *concept*-related components (e.g., "facts" rather than "headings").

Three of the initial editors subsequently participated in a series of exploratory tagging exercises. Each tagging exercise consisted of a preliminary briefing, the tagging exercise itself, and a post-exercise feedback round. In the preliminary briefings, one of three component subsets was addressed. During these meetings, the upcoming tagging exercise was described in detail and the participants were given the opportunity to ask clarifying questions about instances, for example, where it would be possible to interpret aspects of the identification task in more than one way. The tagging exercises themselves involved having the editors identify the component elements and their boundaries in two to three opinions. (Hard copies of the cases were distributed and the participants 'tagged' the components of interest, including their boundaries, using colored pens.) The component subsets used are presented in Section 4.2.2. Finally, during a post-exercise feedback session the editors were presented with the quantitative results of their work in terms of tagging agreements and disagreements (i.e., overlap and misses). During both category-by-category and occurrence-by-occurrence tag reviews, the editors explained their decisions. Procedures and rules were then developed to increase tagging agreements in subsequent exercises.

3.3 Phase III—Reassessment of the Component Set

An identification 'summit' was held with tagging exercise participants. Its objective was to assess each of the components examined during the tagging exercises. Finally, these editors collaborated to establish a modified set of components based on a compendium of quantified results from the tagging exercises as well as upon factors such as utility, cost of identification, and set cohesiveness. This modified set is currently the focal point of on-going experiments.

4. RESULTS

The interview process initially produced a list of more than 150 components. These were consolidated, when appropriate, to collapse duplicate or largely overlapping categories. This resulted in a reduction of the list by one-third: 97 components remained, with partial degrees of overlap still existing between some. The focus group exercise reduced this set to 52 components. They are presented below in Figure 1. These were assembled by their logical groupings, with parallel subset groupings occurring under the topics of *Facts* and *Issues*.

4.1 Phase I

4.1.1 The Resultant Components

See Figure 1.

[3] The terms *component* and *tag* are used interchangeably throughout the paper.

Concept-related Components	**Citations**	**Opinions**
	1. Authority	1. Concurring
Facts	a. Binding	2. Dissenting
1. Historical	b. Non-binding	**Separate Orders**
2. Procedural (See Issues)	2. Statutes	
1. Informative (Embodied above)	3. Cases	**Formatted-related Components**
2. Necessary to Analysis	4. Secondary Sources	
Issues	**Analysis**	**Illustrations**
1. Preliminary	1. Public Policy	**Quotations**
a. Jurisdictional	2. Precedent	— Quoted Transcripts
b. Standard of Review	3. Conclusory Statements	**Paraphrases**
2. Procedural	— Judicial Notice	**Footnotes**
3. Substantive	4. Higher Law	**Appendices**
4. Contentions of Parties	5. Deductive Reasoning	**Headings**
1. Disputed	6. Directed Result	**Divisions**
2. Marginally Disputed	7. Interpretation of Authority	**Judges' Names**
3. Undisputed	8. Application of Law to Facts	**Attorneys' Names**
Law	**Parties**	**Court-supplied Syllabi**
1. Cited (See Citations)	**Evidence**	**Court-supplied Headnotes**
2. **Conclusions**	**Relationships**	**Docket Number**
a. Abstract	**Historical Treatment**	**Date of Decision**
b. Concrete	**Deference to Other Courts**	
Definitions	**The Mandate**	**Sections Not to be Published**

Figure 1. Resultant Components from Data Collection Phase

4.1.2 Typicality and Importance Scoring

Typicality and Importance were two metrics among several which were considered as component selection criteria—i.e., to aid in deciding which to include in subsequent experiments. For this reason, each of the participants was asked to "score" each of the 52 components for *typicality* and *importance*, using a scale from 1 (for most typical/important) to 4 (for least typical/important). The scale was established by the editors themselves as their final focus group exercise. A summary of survey results is presented in Table 1.

4.1.3 Q-sorting Exercise

The Q-sorting exercise produced scores which were extremely consistent with the group mean scores; it also produced a few scores which were quite divergent from the mean. For instance, while one editor had a 0.893 correlation with the group mean, another had a correlation of 0.275. Within the group itself, two editors had a correlation between their scores of 0.758, while another pair had a correlation of 0.009 (one having used his own interpretation of the components rather than the glossary) (See Table 2). It was precisely for instances like the latter that we designed the *Delphi-round* reassessment as a *Q-sort* follow-up exercise.

4.1.4 Delphi-round Reassessment

To avoid individually skewed instances of scoring, the participants were given a second opportunity to assess the typicality and importance of the case components. This time each subject was given a worksheet which included his or her original *Q-sort* scores, along with the group's mean values. Components which were associated with large standard deviations ($s.d. \geq 1.0$) were marked with an asterisk. The objective of the *Delphi-round* was to give the participants the chance to reconsider their original scores in the light of the group mean values (See Table 2).

The *Delphi-round* exercise reduced the weaker correlations discussed above, and as a result, many of the skewed scores converged towards the mean values. The exercise had the net effect of increasing all of the correlation values. (Values shown in Table 2 present Pearson Product-Moment Correlation values.)

4.1.5 Observations on Typicality and Importance

One might expect the editors' typicality scores to be more uniform than those for importance, since frequency (and thus indirectly typicality) is a more mentally recordable item than importance. This was demonstrated in our results.

We speculate that this arises because importance is more subjective than typicality. The figures from both the *Q-sort* and *Delphi-round* exercises support this hypothesis. The mean values of the standard deviations in both Tables 1 and 2 are lower for *typicality*. These figures are shown in Table 3.

Both standard deviation averages decreased following the *Delphi-round* reassessment task, though slightly more for typicality than for importance. These results suggest that after years of closely reading cases, the expert editors develop an internalized notion of common components and their significance to the case. Their notions, however, coincide more for typicality than importance.

The case components were examined after being ordered by typicality and importance scores. It was thought that these orderings could be significant at some point in the study when thresholds are sought to separate useful components from less useful ones. Components which are both typical and important would be the first candidates for frequency analysis during the investigation of actual cases.

The editors were next asked to identify implicit *relationships* between top-level components.[4]

[4]Implicit relationships were viewed as being distinct from the explicit

No.	Component	Mean Typicality Score	Standard Deviation	Mean Importance Score	Standard Deviation
1.	Abstract Conclusions	1.64	1.03	1.91	1.11
2.	Analysis of Precedent	1.45	0.52	1.55	0.52
3.	Appendices	4.00	0.00	3.82	0.40
4.	Appl. of Law to Facts	1.00	0.00	1.27	0.65
5.	Attorneys' Names	1.27	0.90	3.73	0.47
6.	Binding Authority	1.82	0.87	1.45	0.69
7.	Case Citations	1.36	0.67	1.91	0.70
8.	Conclusory Statements	2.09	0.70	2.36	1.03
9.	Concrete Conclusions	1.00	0.00	1.18	0.60
10.	Concurring Opinions	3.09	0.54	3.18	0.87
11.	Contentions of Parties	1.45	0.52	1.91	0.70
12.	Court-supplied Headnotes	3.36	0.67	2.18	1.17
13.	Court-supplied Syllabi	3.45	0.69	2.64	1.21
14.	Date of Decision	1.18	0.60	3.09	0.94
15.	Deductive Reasoning	2.09	0.94	2.09	1.04
16.	Deference	2.73	0.65	2.82	0.40
17.	Definitions	2.64	0.92	2.18	0.87
18.	Directed Result	2.64	1.12	2.55	1.21
19.	Disputed Issues	1.09	0.30	1.36	0.67
20.	Dissenting Opinions	3.09	0.54	3.00	0.89
21.	Divisions	2.36	0.92	3.45	0.93
22.	Docket Number	1.36	0.92	3.64	0.92
23.	Evidence	1.45	0.69	1.82	0.60
24.	Facts Necessary to Analysis	1.09	0.30	1.45	0.69
25.	Footnotes	2.45	1.13	2.73	0.79
26.	Headings	2.91	0.70	3.45	0.69
27.	Higher Law	2.64	1.21	2.09	1.22
28.	Historical Facts	1.55	0.69	2.45	0.93
29.	Illustrations	3.64	0.67	3.36	0.67
30.	Interpretation of Authority	1.73	0.47	1.55	0.52
31.	Judges' Names	1.18	0.60	3.09	1.22
32.	Judicial Notice	3.27	0.65	3.09	0.54
33.	Jurisdictional Issues	2.09	0.54	1.64	0.50
34.	Mandate	1.18	0.60	1.91	1.14
35.	Marginally Disputed Issues	2.82	0.60	2.91	0.94
36.	Non-Binding Authority	2.36	0.92	2.73	0.79
37.	Paraphrases	2.27	0.65	2.73	0.90
38.	Parties	1.00	0.00	2.00	1.00
39.	Preliminary Issues	2.18	0.60	2.73	0.65
40.	Procedural Issues	1.64	0.50	2.09	0.70
41.	Public Policy Analysis	2.91	0.54	2.18	0.75
42.	Quotations	1.55	0.69	2.82	0.75
43.	Relationships	2.36	0.92	2.73	0.90
44.	Secondary Source Citations	2.55	0.52	2.45	0.69
45.	Sections not to be Published	4.00	0.00	3.91	0.30
46.	Separate Orders	3.55	0.93	3.36	1.03
47.	Standard of Review Issues	2.00	0.77	2.18	0.60
48.	Statute Citations	1.64	0.81	1.73	0.79
49.	Substantive Issues	1.09	0.30	1.27	0.65
50.	Transcript Quotations	3.18	0.98	2.82	1.08
51.	Historical Treatment	2.64	0.92	2.18	1.25
52.	Undisputed Issues	2.82	0.60	3.27	0.90
	Mean S.D. Values		0.51		0.70

Table 1. Case Components with Delphi-round *Typicality* and *Importance* Values [1=High; 4=Low]

Correlation Values	*Q-sort*			*Delphi-round*		
	Typicality	*Importance*	Combined	*Typicality*	*Importance*	Combined
Editor–Editor: Min:	0.317	0.000	0.009	0.597	0.046	0.368
Max:	0.797	0.740	0.758	0.876	0.805	0.786
Editor–Group: Min:	0.557	0.096	0.275	0.807	0.531	0.675
Max:	0.896	0.899	0.893	0.911	0.916	0.913

Table 2. Minimum and Maximum Correlation Values for Editor–Editor and Editor–Group Scores

	Typicality	*Importance*
Q-sort	0.64	0.81
Delphi-round	0.51	0.70

Table 3. Mean Values of Composite *Typicality* and *Importance* Standard Deviations from Table 1

4.1.6 Typicality and Importance Relationships

We examined the relationship between typicality and importance scores on a component basis in order to determine if there might be a computational means of measuring importance. If typicality, the less subjective of the two, is in fact related to importance in some tractable manner, then typicality could represent an indirect indicator of component importance. Further, if component *frequency* is a characteristic of typicality, then a researcher may be able to automatically monitor both typicality and importance. A scatter plot of typicality vs. importance scores is shown in Figures 2 and 3. The Pearson correlation score for all 52 components, as shown in Figure 2, is less than 0.6. From the plot, one can observe that the majority of the outliers tend to have low importance but high typicality scores (e.g., *docket number*, *divisions*, *attorneys* and *judges names*). When these format-related components, identified in *Phase I*, are removed from consideration, as shown in Figure 3, the typicality-importance correlation score increases to over 0.84. The suggestion here is that for more significant concept-bearing components, there is a more direct correlation between typicality in an opinion and importance. For instance, *analysis of precedent*, *interpretation of authority*, *application of law to facts*, et al. have strong positive correlations between the two metrics. For any automated or semi-automated system focusing on identifying and indexing these underlying judicial opinion components, such relationships could be meaningful in helping interpret highly developed arguments.

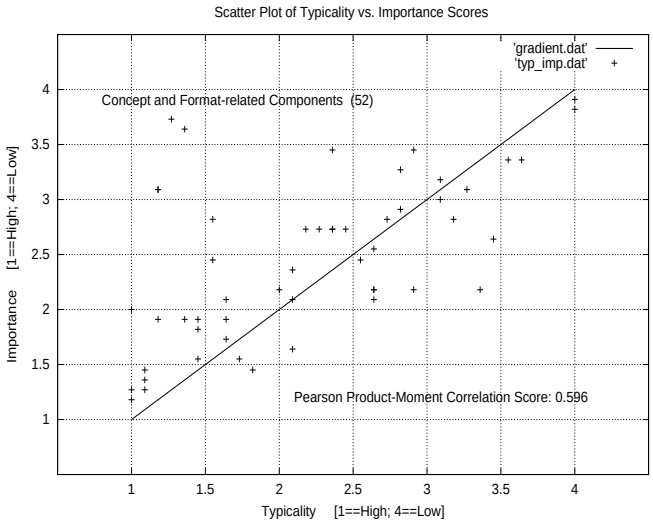

Figure 2. *Scatter Plot of Typicality vs. Importance Scores*

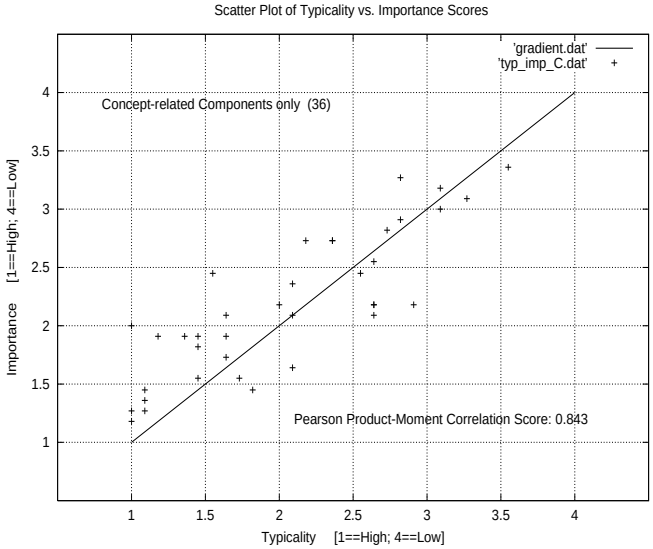

Figure 3. *Scatter Plot of Typ. vs. Imp. Scores (Subset)*

ones—e.g., hierarchical—revealed in Figure 1. The eighteen top-level components were taken from the focus group's initial component outline (boldfaced entries in Figure 1). A relationships questionnaire asked editors to "denote (with symbols and notes) the types of relationships you believe may exist between the individual component and any of the others." A distinction was made between conceptual and format-related topics. The task was made somewhat more manageable for the editors because no components which are already 'tagged' in the opinion have been included (e.g., judges' names, decision date, docket number, etc). Results from the questionnaire were recorded in a knowledge-base organized by (a) component and (b) link. We envision drawing from this repository at a point when seeking discriminating component characteristics. Because the nature and range of this information is beyond the scope of this paper, it will not be further addressed here.

4.2 Phase II

4.2.1 The Tagging Exercise Component Set

The second phase of the study began by having each of the initial subjects participate in a tagging survey. Tagging in this context was defined as the act of inserting tags in a textual document to identify a particular element or feature (i.e., component) and marking its boundaries. Different tags could thus be characterized by different scopes (i.e., some would occur at the word-level, others at the phrase or sentence-level, and still others at the paragraph or multiple paragraph-level). In the survey, participants were asked to respond to two separate tag-related indicators—component identification and boundary designation—as well as a hybrid of the two. The results of the hybrid category follow. The content-(i.e., non-format)-related component set used in five tagging exercises is presented in Figure 4, grouped by tagging survey results. The survey assessed each component according to presumed tagging difficulty; the components were initially grouped into three subsets according to this relative tagging difficulty. [The figures in parentheses give each component's typicality (T) and importance (I) *rank* among the group of 52 components.]

The 36 components shown in Figure 4 below were subsequently subdivided into three different groups based largely, though not exclusively, upon their tagging survey scores.[5] Each group included components from the other two scoring groups. To some extent, each subgroup focused on a topical area. For example, the second group was designed to focus as much as possible upon *issue*-related components. The resultant regroupings appear in Figure 5.

4.2.2 Tagging Exercise Results

Throughout our analysis, tagging exercise discoveries were as much qualitative as quantitative; they arose from discussions following the exercises as well as from our numerical spreadsheets. Nonetheless, tagging concordances among the editors were routinely monitored in a quantitative manner. After experimenting with several different levels of granularity, we finally selected the paragraph-level as a reasonable indicator of tagging agreement. For each of these specified segments of text, the degree of agreement among editor pairs and editor trios was recorded. These concordances tended to correspond to the established consensus about which components were more difficult to identify and which less so. Average concordances from three of the five tagging exercises are presented in Table 4. For cells in the table reporting two values, the first value includes those paragraphs in which the editors agreed contained *no* tags; the second value excludes them. Sections of a test case where all three editors were in agreement on the tags were rarer, especially for the more challenging tags sets. All of the tagging exercises were performed on federal court opinions. With the exception of an initial meeting to discuss definitions for a given tag set and the specific task asked of them, the editors performed their tagging independently of one another. The first tagging exercise with component subset I was regarded as a training round. The third tagging exercise (also on subset I) was performed using different test cases and thus do not appear in the table. (Thus, a set of summary results from the remaining tagging exercises appears below.) As the editors moved from the easier tag set to the more difficult one, overall agreements about what portions of the text should be tagged declined (as should have been expected).

5. CONCLUSIONS

One of the chief observations made during our component analysis involved the cost-quality tradeoff. In other words, the more extensive the tag set, the more costly the investment required to identify them reliably, whether manually or automatically. This finding was especially true of the more challenging components in the last two tagging subsets. Difficulties encountered in identifying the components were many. Some of the more basic reasons included the broad range covered by a given tag (e.g., *Relationships*), the slippery nature of a tag's definition (e.g., *Directed Result*), the questionable contribution of a tag (e.g., *Non-binding Authority* and *Marginally Disputed Issues*), the infrequency of a tag (e.g., *Higher Law*), the low degree of overlap among editors when identifying a given tag (e.g., *Paraphrases* and *Deductive Reasoning*), equal representation by other existing tags (e.g., *Deference* by *Binding Authority* or *Historical Treatment*), and combinations of the above. In addition, when the editors were in reasonable agreement on a given tag, the demarcation of the tag's boundaries presented an entirely different challenge, particularly for general tags such as *Analysis of Precedent* or *Application of Law to Facts* where the component can span over multiple paragraphs.

Beyond the particular sources of difficulty mentioned above, one concern was raised on several occasions by the participants. Given our component set and its three subsets, the tags were seen to lack a unifying view or coherent backbone which could tie together important elements such as issues and corresponding evidence, analysis, and conclusions. This apparent disunity was explicitly addressed in Phase III.

Focusing on issues-related components, the participants began to underscore probable problem areas during our discussion sessions. It was believed that a potentially important class of components was being unduly diluted because of a mixture of precise tags with more general ones. A combination of Typicality and Importance scores from Phase I was also taken into account in this portion of the experiment. For example, from a practical point of view *Preliminary Issues* was seen as an extraneous tag, since virtually all issues would fall under the classes of *Procedural Issues* and *Substantive Issues*. So such a tag would rarely, in practice, be required. Through a series of such reductions, an initial class of ten issues-related components could be more succinctly represented by five.

Similar arguments were made for a number of fact-related components, namely, *Historical Facts*, *Judicial Notice*, *Facts Necessary to Analysis*, as well as *Evidence*. In order to properly capture the nature and role of evidence in a case, we questioned whether these could be more clearly represented with fewer elements, but more essential ones. The expert editors agreed they could. As a result, a subsequent goal became taking a step backwards from the initially diffuse tag set to "distinguish the forest from the trees." The underlying motivation was that if agreement could not be achieved using manually-assigned tags on a training set, then little success could be expected when tagging these components automatically.

[5]The *Conclusory Statements* component was further clarified by and for the participants during the construction of the tagging glossary. At that time it was more explicitly described as *Summary Resolution of an Issue*.

Mean Score Results of Tagging Survey (Responses from 10 Participants)

Relatively Easy [1.0, 1.5]	Scores	Moderately Easy]1.5, 2.25]	Scores	Relatively Difficult]2.25, 4.0]	Scores
The Mandate (T: 1; I: 15)	1	Contentions of Parties (T: 14; I: 13)	1.60	Directed Result (T: 33; I: 29)	2.45
Separate Orders (T: 49; I: 47)	1	Preliminary Issues (T: 28; I: 31)	1.65	Analysis of Precedent (T: 15; I: 6)	2.55
Concurring Opinions (T: 43; I: 43)	1	Jurisdictional Issues (T: 27; I: 10)	1.80	Public Policy Analysis (T: 42; I: 18)	2.60
Dissenting Opinions (T: 43; I: 42)	1	Historical Facts (T: 16; I: 22)	1.80	Application of Law to Facts (T: 1; I: 2)	2.65
Transcript Quotations (T: 45; I: 37)	1	Higher Law (T: 38; I: 22)	1.80	Facts Necessary to Analysis (T: 8; I: 4)	2.75
Judicial Notice (T: 45; I: 38)	1.10	Disputed Issues (T: 8; I: 4)	1.85	Deductive Reasoning (T: 23; I: 13)	3.05
Undisputed Issues (T: 38; I: 45)	1.10	Marginally Disputed Issues (T: 38; I: 38)	1.90	Relationships (T: 25; I: 28)	3.80
Standard of Review Issues (T: 23; I: 22)	1.20	Substantive Issues (T: 8; I: 3)	1.95		
Second. Source Citations (T: 34; I: 27)	1.20	Non-Binding Authority (T: 30; I: 36)	2		
Definitions (T: 36; I: 21)	1.40	Treatments (T: 35; I: 15)	2		
Binding Authority (T: 21; I: 6)	1.40	Procedural Issues (T: 18; I: 18)	2.05		
Sum. Resolution of Issue (T: 25; I: 26)	1.50	Interpretations of Authority (T: 21; I: 8)	2.10		
		Paraphrases (T: 28; I: 31)	2.10		
		Abstract Conclusions (T: 18; I: 15)	2.10		
		Concrete Conclusions (T: 1; I: 1)	2.15		
		Evidence (T: 13; I: 10)	2.20		
		Deference (T: 36; I: 34)	2.20		

Scoring: 1=Quite Easy; 2=Moderately Easy; 3=Moderately Difficult; 4= Quite Difficult.

Figure 4. *Tagging Survey Resultant Groupings*

Component Subsets Used in Tagging Exercises

Subset I	Subset II	Subset III
The Mandate	Contentions of Parties	Standard of Review Issues
Separate Orders	Preliminary Issues	Historical Treatment
Concurring Opinions	Jurisdictional Issues	Interpretations of Authority
Dissenting Opinions	Sum. Resolution of Issue	Abstract Conclusions
Transcript Quotations	Undisputed Issues	Concrete Conclusions
Judicial Notice	Disputed Issues	Evidence
Second. Source Citations	Marginally Disputed Issues	Deference
Definitions	Substantive Issues	Analysis of Precedent
Binding Authority	Procedural Issues	Public Policy Analysis
Historical Facts	Deductive Reasoning	Application of Law to Facts
Higher Law	Paraphrases	
Non-Binding Authority		
Directed Result		
Facts Necessary to Analysis		

Figure 5. *Tagging Exercise Component Subsets*

Tagged Opinion	Paragraphs in the Opinion	Subset I		Subset II		Subset III	
		2 Eds.	3 Eds.	2 Eds.	3 Eds.	2 Eds.	3 Eds.
716 F.2d 415	17	71%	24%	74%/16%	16%/0%	42%/32%	53%/11%
716 F.2d 234	44	93%	59%	88%/26%	41%/0%	60%/38%	17%/ 2%

Table 4. *Tagging Exercise Concordances Among Editors*

After having focused on identifying components in the set, we proceeded to examine them using the dominating issues as an organizing principle. Whereas issues initially represented something to be resolved, they are herein permitted to acquire a dual-role, their second function being a thread to which other components become attached. The advantages of such dual functionality include an organizing structure that has natural links to existing topical views of a case (e.g., bankruptcy, immigration, insurance, etc).

Modifications of the Component Model (Phase III)

In Phase III, each of the participants was asked to draft a judicial opinion component set of his own. The consensus produced was that in the interest of utility these sets would be less fine-grained than those addressed in the tagging exercises. In the resultant versions, many of the higher-level members of our original set reappeared, but as part of broader categories. For this reason, the editors largely agreed on a core set of components. Their selection criteria shared three significant features, namely, utility, cohesion, and cost to delimit, in addition to typicality and importance. The results are presented below, along with insights about the identification process and the role these case components might play in editorially-enhanced legal opinions.

The synthesis of components appearing in the revised set includes:

1. **Main Set**—characterized by their ability to be linked to fixed issues or annotations.
   (a) *Issues/Contentions*
      i. *Substantive*
      ii. *Procedural*
   (b) *Analysis* (with subcomponents such as *Policy Analysis* preserved for future use.)
   (c) *Facts/Evidence*
   (d) *Conclusions* (including *Mandate*)
      i. *Abstract*
      ii. *Concrete*
2. **Subsidiary Set**—useful tags despite an absence of connections to those in the main branch.
   (a) *Binding Authority*—cases that compel a court to come to the conclusions it does, more precisely referred to as "Principal Authority" or "Controlling Authority."
   (b) *Historical Treatment*—especially including "the court's language criticizing the cited case."
   (c) *Definitions*—"explanations of meaning of a word, phrase, or context."
   (d) *Other Components*—which are effectively tagged by virtue of current format conventions, e.g., statutes and case citations, concurring and dissenting opinions, and other separate opinions and separate orders.

The notion of a core component set and a subsidiary set was proposed by a number of the participants. The rationale was that members of the core set could be linked to an editorial view or other standardized heading (e.g., Conclusions could be tied to a given Issue while Facts/Evidence and Analysis could be tied to the Issues and Conclusions). This main set is thus the most cohesive. The subsidiary set may have less cohesiveness but nonetheless merits tagging for its value to the legal researcher.

The idea of an expanding tag set was also supported, or at least one evolving from the first level of generality to the second or sub-component level. This would avoid the need to have a special training course while, for example, attempting to implement all candidate tags at once.

The results of our study are promising, though questions of scope may arise from the limited size of our tagging data sets. Our findings suggest that a closer examination of the supra-textual, discourse-level components of full-text judicial opinions reveals features which have traditionally remained untapped. Moreover, the potential of this research to assist with the goal of the semantic tagging of legal text may be its most valuable contribution. Although it is unlikely that these components will lead to the replacement of statistical word-based retrieval systems any time soon, the potential for such components to enhance existing access methods—especially in the area of case law—is encouraging. And many would agree that there is much room for improvement in current natural language systems.

6. FUTURE WORK

In order to demonstrate that our distilled set of case components can be used to establish a richer and more coherent set of accessible features, additional experiments were called for. Our findings have supported the use of issues as an organizing principle around which the other related components are set in motion. Once editors and ultimately computational resources are trained to manage identification of such a central feature, work can advance to the next logical tags, tags such as evidence, analysis, and conclusions. We are currently examining such a network of components through the use of training and test case sets. We have not previously exploited the relationships discovered between the components. For this reason, we are empirically examining some of the various links which exist between the components found in standard judicial opinions. Once they have been manually identified in a set of training cases, we will

proceed to further characterize these features in additional experiments.

7. ACKNOWLEDGEMENTS

We thank Frances Lawrenz from the University of Minnesota for her helpful suggestions concerning our methodology. We are grateful to Howard Turtle for his continued support and useful comments during every phase of this project. We wish to thank Charles Swope and Brian Johnson for granting us the extensive editorial resources which were required for our experiments. This work also benefited from on-going conversations with a number of editorial colleagues, especially Andrews Allen Jr., Andrew Martens, Vincent Platt, and Steven Sweeney.

8. REFERENCES

[1] R. J. Aldisert. *Opinion Writing*. West Publishing, St. Paul, 1990.

[2] V. Aleven. *Teaching Case-Based Reasoning Argumentation through a Model and Examples*. Ph.d. dissertation., University of Pittsburgh, 1997.

[3] S. J. Alvarado. *Understanding Editorial Text: A Computer Model of Argument Comprehension*. Kluwer Academic Publishers, Boston, 1990.

[4] W. R. Borg and M. D. Gall. *Educational Research: An Encyclopedia of Research Methods and Techniques, 4th ed.* Longman, White Plains, NY, 1983.

[5] S. Brüninghaus and K. D. Ashley. Toward adding knowledge to learning algorithms for indexing legal cases. In *Proceedings of the 7th Int'l Conference on Artificial Intelligence and Law*, pages 9–17. ACM ICAIL, 1999.

[6] B. K. Britton and E. John B. Black. *Understanding Expository Text: A Theoretical and Practical Handbook for Analyzing Explanatory Text*. Lawrence Erlbaum Associates, Hillsdale, NJ, 1985.

[7] R. Cohen. Analyzing the structure of argumentative discourse. *Computational Linguistics*, 13(1-2):11–24, Jan-June 1987.

[8] S. Coval and J. Smith. *Law and Its Presuppositions*. Routledge & Kegan Paul, London, 1986.

[9] B. Danet. Language in the legal process. *Law & Society Review*, 14(3):455–564, Spring 1980.

[10] J. J. Daniels and E. L. Rissland. A case-based approach to intelligent information retrieval. In *Proceedings of the 18th Annual Int'l ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 239–245. ACM SIGIR, 1995.

[11] C. Deedman and J. Smith. The nervous shock advisor: A legal expert system in case-based law. In C. Y. Suen and E. Rajjan Shinghal, editors, *Operational Expert System Applications in Canada*. Pergamon Press, New York, 1991.

[12] D. Gelbart and J. Smith. Automated paragraphs extraction from legal documents. Technical report, University of British Columbia, Faculty of Law and Artificial Intelligence Research (*FLAIR*), 13 pages, Vancouver, Canada, 1993.

[13] M. A. Hearst and C. Plaunt. Subtopic structuring for full-length document access. In *Proceedings of the 16th Annual Int'l ACM SIGIR Conference on Research and Development in Information Retrieval (Pittsburgh, PA)*, pages 59–68. ACM SIGIR, June-July 1993.

[14] L. B. Jones. *Pragmatic Aspects of English Text Structure*. The Summer Institute of Linguistics and The University of Texas at Arlington, 1983.

[15] F. Lawrenz. *Personal Communication*. Office of the Vice-President of Research, College of Education and Human Development, University of Minnesota, Minneapolis, MN, Oct 1994.

[16] E. D. Liddy. The discourse-level structure of empirical abstracts: An exploratory study. *Information Processing & Management*, 27(1):55–81, 1991.

[17] R. E. Longacre. *The Grammar of Discourse*. Plenum Press, New York, 1983.

[18] W. C. Mann and S. A. Thompson. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281, 1988.

[19] D. Marcu. Discourse trees are good indicators of importance in text. In I. Mani and M. T. Maybury, editors, *Advances in Automatic Text Summarization*, pages 123–136. MIT Press, Cambridge, MA, 1999.

[20] D. Mellinkoff. *The Language of Law*. Little, Brown, Boston, 1963.

[21] C. D. Paice and P. A. Jones. The identification of important concepts in highly structured technical papers. In *Proceedings of the 16th Annual Int'l ACM SIGIR Conference on Research and Development in Information Retrieval (Pittsburgh, PA)*, pages 69–78. ACM SIGIR, June-July 1993.

[22] D. E. Rumelhart. Schemata: The building blocks of cognition. In R. J. Spiro, B. C. Bruce, and E. William F. Brewer, editors, *Theoretical Issues in Reading Comprehension: Perspectives from Cognitive Psychology, Linguistics, Artificial Intelligence and Education*, pages 33–58. Lawrence Erlbaum Associates, Hillsdale, NJ, 1980.

[23] W. Stephenson. *The Study of Behavior: Q-Sort Technique and Its Methodology*. University of Chicago Press, Chicago, IL, 1953.

[24] R. Susskind. *Expert Systems in Law*. Clarendon Press, London, 1987.

[25] H. R. Tibbo. *Abstracts, Online Searching, and the Humanities: An Analysis of the Structure and Content of Abstracts of Historical Discourse*. Ph.d. dissertation., University of Maryland College Park, 1989.

[26] B. F. Udinsky, S. J. Osterlind, and E. Samuel W. Lynch, editors. *Evaluation Resource Book: Gathering, Analyzing, Reporting Data*. EDITS Publishers, San Diego, CA, 1981.

[27] T. A. van Dijk. *Macrostructures: An Interdisciplinary Study of Global Structures in Discourse, Interaction, and Cognition*. Lawrence Erlbaum Associates, Hillsdale, NJ, 1980.

[28] T. A. van Dijk. *Handbook of Discourse Analysis*. Academic Press, New York, 1985.

[29] T. A. van Dijk. *News as Discourse*. Lawrence Erlbaum Associates, Hillsdale, NJ, 1988.

[30] B. R. Witkin. Future methods: Forecasts, scenarios, and the delphi method. In *Assessing Needs in Educational and Social Programs*, pages 33–58. Jossey-Bass, San Francisco, CA, 1980.

9. APPENDIX

Below is the glossary of case components (abridged). They are customarily not tagged in opinions.

1. **Contentions of Parties**—The dispute; the subject of litigation; the matter for which the suit has been brought.
2. **Preliminary Issues**—Those matters relating to the judge's necessary determination of why the evidence at hand merits a court hearing.
3. **Substantive Issues**—Those matters relating to the part of the law which creates, defines, and regulates the rights and duties of the parties.
4. **Undisputed Issues**—Those matters relating to the case about which the parties are in agreement.
5. **Marginally Disputed Issues**—Those matters relating to the case about which the parties are in partial disagreement.
6. **Disputed Issues**—Those matters relating to the case about which the parties are largely in disagreement.
7. **Procedural Issues**—Matters relating to the legal methods of enforcing rights or obtaining redress for their violation.
8. **Jurisdictional Issues**—Matters relating to the powers of the courts to inquire into facts, apply the law, make decisions, and declare judgments; they address the legal right by which judges are permitted to exercise their authority.
9. **Interpretations of Authority**—A part of the analysis; the means by which the court concludes its scope of control in the litigation at hand, its right to exercise power and to implement and enforce the law as well as the role the judgments of other parallel or higher courts should have in its mandate.
10. **Higher Law**—A part of the analysis; a consideration of the decisions made by courts higher than in the current jurisdiction.
11. **Binding Authority**—Sources of law that must be taken into account by a judge in deciding a case; for example, statutes or decisions by a higher court of the same state on point.
12. **Non-Binding Authority**—Sources of law that need not necessarily be taken into account by a judge in deciding a case; for example, decisions by a lower court or parallel courts in a different jurisdiction.
13. **Deference**—The respect which needs to be paid in decision-making to higher court decisions as well as sources of law which come from binding authorities.
14. **Historical Treatment**—References to prior cases which have affected their precedent-holding value.
15. **Standard of Review Issues**—Those topics which may serve to limit the scope of the court's considerations during a case, e.g., the domain of the court's authority, issues relating to statutes of limitations, jurisdictional issues, etc.
16. **Historical Facts**—Facts from the past which relate to the case at hand and which may or may not be crucial to the resolution of the claims, e.g., the age of the parties or witnesses or evidence, the condition of the environment during the period the claim addresses, etc.
17. **Facts Necessary to Analysis**—Actual and absolute realities to be considered by the court, as distinguished from fiction or error.
18. **Evidence**—Any species of proof, or probative matter, legally presented at the trial of an issue, by the act of the parties and through the medium of witnesses, records, documents, exhibits, concrete objects, etc., for the purpose of inducing belief in the minds of the court or jury as to their contention.
19. **Application of Law to Facts**—A part of the analysis; the means by which the court associates the evidence at hand with the law in order to make a judgment.
20. **Secondary Source Citations**—References to sources other than statutes or previously judged cases; they may include Legal Reviews, textbooks, periodicals, articles, letters, notices, and many other printed or copyrighted materials.
21. **Transcript Quotations**—A verbatim reference to a copy of the trial, hearing, or other proceeding as prepared by the court reporter.
22. **Analysis of Precedent**—A form of analysis which involves an examination of prior court rulings which are close in facts or legal principles to the case under consideration.
23. **Public Policy Analysis**—A part of the analysis which examines the suitability of existing community common sense prescriptions and enactments.
24. **Deductive Reasoning**—A form of analysis; the process of deriving a result through the use of inference in which the conclusion follows necessarily from the premises.
25. **Directed Result**—A form of analysis which uses particular and focused logical reasoning to arrive at its conclusion.
26. **Conclusory Statements (or Summary Resolution of an Issue)**—Statements, distinct from judgments or final conclusions which summarize the result of directed analysis or reasoning thus far.
27. **Judicial Notice**—A type of conclusory statement; the act by which a court, in conducting a trial, or framing its decision, will, of its own motion or on request of a party, and without the production of evidence, recognize the existence and truth of certain facts, having a bearing on the case, which are not properly the subject of testimony or which are universally established (e.g., laws of the state, international law, historical events, the constitution, etc).
28. **Abstract Conclusions**—Determinations, dissassociated from any specific instance, which draw to a close a line of argument or reasoning, e.g., that segregated schools are inherently unequal.
29. **Concrete Conlusions**—Deteminations, based on specific instances, which draw to a close a line of argument or reasoning, e.g., that Ms. Brown was discriminated against by the Topeka schools and she will be permitted to attend Topeka High.
30. **The Mandate**—A command, order, or direction, written or oral, which the court is authorized to give and the parties are legally bound to obey.
31. **Separate Orders**—Directions of a court or judge made or entered in writing, and not included in a judgment which determines some point or directs some action following the proceedings.
32. **Concurring Opinions**—A statement of judicial conclusion which agrees with the result reached by the majority, but disagrees with the precise reasoning that lead to that result.
33. **Dissenting Opinions**—A statement of judicial conclusion which disagrees with the result reached by the majority and thus disagrees with the reasoning and/or principles of law used by the majority in deciding the case.