"Steps" towards a corpus of SCOTUS opinions annotated using a Swalesian approach

Warren Bonnard[1], Mary Catherine Lavissiere[2], Anas Belfathi[2], Nicolas Hernandez[2], Christine Jacquin[2] & Laura Monceaux-Cachard[2]
University of Lorraine (France)[1], Nantes Université (France)[2]
warren.bonnard@univ-lorraine.fr, marycatherine.lavissiere@univ-nantes.fr, anas.belfathi@univ-nantes.fr, nicolas.hernandez@univ-nantes.fr, christine.jacquin@univ-nantes.fr, laura.monceaux@univ-nantes.fr

Abstract

In the tradition of Moreno and Swales (2018), this paper presents the creation of a manually annotated resource for supporting teaching English for Legal Purposes (ELP) and for Natural Language Processing (NLP) purposes. After justifying the use of Supreme Court of the United States opinions, we define our coding scheme by adapting the move model of rhetorical structure in specialized discourse. We describe the methodology and the implementation of the annotation campaign. We analyze how our methodology and the resulting annotation scheme diverge from those described in the literature as well as the advantages that these divergences afford. In addition to the research article, we release several supplementary materials which aim to make the process transparent and serve other researchers aiming to annotate specialized discourse with the help of machine learning techniques.

Keywords: Move analysis, discourse analysis, case law, English for Legal Purposes, annotation.

Résumé

«Steps» vers un corpus d'avis de la SCOTUS annotés selon l'approche Swalesienne

Dans la lignée de Moreno et Swales (2018), cet article présente la création d'une ressource annotée manuellement pour soutenir l'enseignement de l'anglais juridique et le traitement automatique des langues. Après avoir justifié l'utilisation

des opinions la Cour suprême des États-Unis, nous définissons notre schéma d'annotation en adaptant le modèle de Swales (2004) au discours juridique. Nous décrivons la méthodologie et la mise en œuvre de la campagne d'annotation. Nous analysons les contributions spécifiques de notre méthodologie et du schéma d'annotation correspondant par rapport aux cadres méthodologiques existants et décrits dans la littérature. En parallèle de l'article de recherche, nous publions plusieurs documents supplémentaires qui visent à accroitre la transparence de notre méthodologie. Ces documents pourront être utiles à d'autres chercheurs souhaitant annoter des discours spécialisés avec le concours de techniques informatiques d'apprentissage automatique.

Mots-clés: analyse swalesienne, analyse du discours, arrêts, anglais juridique, annotation.

1. Introduction

Judicial opinions written by judges characterize common-law legal systems. When binding precedent in common-law countries, these documents have the same weight as law created through the legislative process. This is especially true for the United States. The Supreme Court in the United States (SCOTUS), for example, can strike down Acts of Congress. Case law also influences civil law countries, especially in situations where both civil and common law coexist, such as the case in Canada and until the United Kingdom's exit, also in the European Union (EU).

In educational contexts, law students read many legal opinions during their training. In language learning contexts, authentic judicial opinions are difficult for English for Legal Purposes (ELP) learners to read (Lavissière et al., 2024) and challenging for ELP teachers to integrate into their courses (Boulton et al., 2025). This difficulty arises for many reasons. The language of legal opinions is complex (Kirby-Légier, 2005); it mostly lacks section headings which would allow for easier reader navigation (Yaich & Hernandez, 2025); the opinions are often long and may be seen as monotonous for the ELP learner. ELP teachers may also find them intimidating. The literature on ELP in France, for example, shows that these traits may reduce the use of actual judicial opinions or extracts in ELP courses (Boulton et al., 2025).

This study is part of a larger national project, Lexhnology, that aims to facilitate access to case law for ELP teachers and learners. Lexhnology represents a collaboration between language teachers, linguists, and Natural

Language Processing (NLP) researchers. The main assumption of the project is that access to judicial opinions annotated with moves and/or steps (Swales, 1990, 2004) through a corpus exploration tool will improve ELP reading comprehension. However, NLP methods for annotation of discourse units with information about the communicative acts of functions, as in move analysis, have not yet been applied to American case law (for applicable to scientific articles, see Teufel et al., 1999; for Indian legal texts, see Kalamkar et al., 2022).

A first step towards facilitation of reading case law is the description of its generic structure, here through move analysis (Swales, 1990), one of the major theoretical frameworks for teaching and learning English for Specific Purposes (ESP). Following the methodological framework described by Moreno and Swales (2018), this paper aims to answer two research questions:

> 1) How can move analysis methodology be adapted for use in machine learning?
>
> 2) How can move analysis methodology be adapted for use in legal discourse?

This paper is structured as follows: In Section 2, move analysis and its applications to legal texts are presented. Section 3 describes the preparatory work for the annotation campaign, including corpus compilation, and the creation of the coding scheme and annotation guide. In Section 4, the annotation procedure and the measures implemented to verify the robustness of the proposed resources are detailed. The method used and the results are discussed considering the literature in Section 5. Conclusion and perspectives for future work are given in Section 6.

In addition to discussing methodological issues, this paper includes the publication of a number of supplementary materials. These show the actual products of the research and may be useful to the community for understanding legal texts or texts in other specialized domains. They include the coding scheme (in English), the Label Studio (Tkachenko et al., 2020) template (in French), the annotation guide (in French), and the annotator training program (video, decision tree, quizzes in French) and parser. The annotated dataset will be released at the end of the project in open source.

2. Move analysis

This section contextualizes the theoretical and methodological and theoretical framework of the present study as regards existing literature.

2.1. ELP orientation of the Lexhnology project

Lexhnology is oriented towards understanding, teaching, and learning ELP. In the absence of top-down information processing guidance for reading in the ELP classroom, the Lexhnology research team undertook the creation of a pedagogical online tool for improving reading comprehension of judicial opinions. This tool will allow teachers and students to access and explore a corpus of majority opinions of SCOTUS that are annotated using the Swalesian model of discourse units. To efficiently annotate a large number of documents, we employed automatic discourse structuring methods from NLP.

While there are many theoretical frameworks available to describe discourse structures in the field of NLP (see Rhetorical Structure Theory (RST) in Mann & Thompson, 1988; or Segmented Discourse Representation Theory (SDRT) in Asher & Lascarides, 2003), their pedagogical contributions to the field of ESP remain limited. In contrast, rhetorical move-step analysis is a primary method to reveal the rhetorical structure of specialized texts in the tradition of the genre analysis ESP framework (Dudley-Evans, 2002). Considering the pedagogical goals of Lexhnology for ELP, this paper has adopted the move analysis framework.

2.2. Rhetorical move-step analysis

Move analysis framework was constructed by John Swales in the context of teaching academic writing in English to non-native researchers. The CARS model (Swales, 1990) divides research article introductions into "moves," which are abstract units that serve one rhetorical purpose (Swales, 1990, 2004) and "steps", smaller concrete units of language that carry out and compose the move (Biber et al., 2007).

The Swalesian framework has been widely applied for modelling specialized language for teaching and learning purposes, especially in the field of English for Academic Purposes to teach writing skills (Yang et al., 2023). Studied academic genres include entire research articles (Cotos et al., 2015); sections of research articles (Le & Pham, 2020); paper abstracts (Salager-Meyer, 1992); PhD dissertations (Soler Monreal, 2016).

This suggests that move analysis is particularly suited for well-studied and codified genres, such as research articles. Research articles generally follow a common rhetorical organization, which can vary in specific ways across different disciplines and subdisciplines. As a result, move analysis frameworks for these texts are often constructed incrementally, adapting existing frameworks "to account for the rhetorical activity of the community practice in focus" (Casal & Kessler, 2023, p. 86).

2.3. Legal English and ELP

In recent years, a wide range of professional legal genres have been subjected to move analysis, offering insights into their rhetorical structure and practical applications for ELP learners. Studies have examined genres from both the domains of public and private law: law abstracts (Breeze, 2009), press releases (Tessuto, 2021), or patent specifications (Groom & Grieve, 2019). Corpus-based analyses of case law —judicial decisions serving as legal precedents— have covered many legal cultures: Poland (Gozdz-Roszkowski, 2020), the United States (Lavissière & Bonnard, 2024) the United Kingdom (Bhatia, 1993; Mazzi, 2007), or China (Han, 2011). These studies account for the discourse specificities related to each national legal context, with frameworks primarily built on corpus data rather than existing frameworks, except for Mazzi (2007), who expanded upon Bhatia's (1993) framework. Drawing on these distinct analyses, Han et al. (2018) identified rhetorical features universal to the genre of the judicial opinion, that is, a "common coercive nature and a shared unchangeable deductive format of legal reasoning" (p. 464).

In each of these analyses of court decisions, the structure can be divided into large "bulky sections" (Han et al., 2018, p. 465) that can be summarized as Facts, Issues, Arguments, and Conclusions/Decision. Interestingly, these findings align with the recommendations issued by US professionals on opinion writing (see Lavissière & Bonnard, 2024), and also with the central moves of press releases from the European Court of Justice (ECJ), which are intended to mirror the structure of ECJ judgments (Tessuto, 2021).

However, the move analysis studies on case law highlight two problems: 1) Little is known about the structure of US judicial opinions from corpus-based research, and 2) the resources written by researchers, lawyers, judges, and jurists do not give fine-grained information about the rhetorical structure of opinions, especially on their core justification, that is, the

presentation of arguments to justify the final decision. On this last point, two exceptions stand out in the literature. On one hand, Gozdz-Roszkowski (2020) found that the legal justifications of Polish Constitutional Judgments involved five major moves, occurring in a specific and fixed sequence. On the other hand, Mazzi (2007) identified four sub-moves to the move *Arguing the case* previously identified by Bhatia (1993): in common law cases, the sub-moves are *Stating history of the case*, *Identifying the conflict of categorisation*, *Presenting Arguments* and *Deriving ratio decidendi;* in ECJ cases, the two sub-moves are *Arguments of the parties* and *Arguments of the court,* showing that the justification of the case is based on the interplay of different judicial voices.

2.4. Methodological framework for move analysis

In constructing specific move-step frameworks, move analysts often adopt a combination of top-down and bottom-up approaches (see Biber et al., 2007, for a complete description of the stages involved in each approach). These approaches "either (a) prioritize steps or linguistic devices but eventually proceeds to the move level or (b) integrate both content/propositional, discoursal, and linguistic cues to carry out the analysis" (Kim et al., 2024, p. 3). Examples of such practices can be seen in studies by Cotos et al. (2017), Moreno and Swales (2018), and Le and Pham (2020).

However, theoretical recommendations on performing move analysis (i.e., selecting an approach) do not provide clear guidance on how to apply the stages in practice. This has resulted in a lack of uniformity in the reporting of methodological practices, especially in areas such as annotation procedures, framework development, unit segmentation, and disagreement resolution (Casal & Kessler, 2023, p. 86). To address this variability, recent move analyses have provided more detailed and systematic accounts of move analysis practices, including information about annotators, annotating protocols, and methods for quantifying agreement between annotators (e.g., interrater reliability scores). For instance, Moreno and Swales (2018) insist on the importance of seeking feedback from experts, checking intra-rater consistency and inter-rater reliability in the protocol used to annotate research articles for steps. According to Kim et al. (2024), this trend reflects a broader effort to strengthen methodological procedures, aligning with a general movement toward greater rigor in applied linguistics research practices (see, for example, Larsson et al., 2023).

2.5. Reliance on NLP models to enlarge the scope of move analysis

Discourse analysis has been described as "a fundamental problem in the ACL (Association for Computational Linguistics) community, where the focus is to develop tools to automatically model language phenomena that go beyond the individual sentences" (Joty et al., 2019, p. 12). Automatic tools, built with computational models, are therefore promising for genre analyses as they offer a way to address the methodological challenges and time-intensive nature of manual coding for moves and steps. However, the use of large language models (LLMs) to enhance move analysis is a recent development, and its application has so far been limited to highly studied and codified genres, such as research article abstracts, which involve a restricted number of move labels (Yu et al., 2024).

Despite its potential, the use of machine learning techniques in this domain remains rare compared to manual annotation in move analysis studies. For instance, Teufel et al., (1999) modified Swales' annotation scheme to automatically detect rhetorical moves in scientific articles. Anthony and Lashkia (2003) developed a computer tool capable of automatically identifying moves in research articles with an overall 68% accuracy rate. In the legal domain, Kalamkar et al. (2022) adopted a framework similar to move analysis to automatically annotate the discursive units of full Indian judicial decisions with 12 "rhetorical roles". These labels are founded on legal concepts, such as precedent, facts, procedural history, and holding.

2.6. Framework for the present study

None of the methodological frameworks presented in the aforementioned studies are entirely satisfactory for the purposes of the present study. The most rigorous reports on move analysis procedures do not focus on their integration to NLP-based move analysis. Conversely, NLP-oriented move analyses predominantly describe the machine learning techniques used to build models, without addressing how these procedures should be adapted for the automatic investigation of understudied genre structures.

Regarding Kalamkar et al. (2022), the complex and subtle construction of arguments in legal opinions was not sufficiently taken into account by the coding scheme proposed in their study. This scheme does not allow for fine-grained annotation of the way in which some judges build their reasoning, especially in the analysis section of the opinion. In contrast, SCOTUS opinions are argumentative prose and as such, a priority in the research

presented here was to focus on the justices' way of constructing and supporting their arguments with legitimate sources of authority. It also became clear that the justices orchestrate and compare many different "voices" during the construction of their arguments, especially in their analysis. For this reason, the coding scheme was also based on dialogic linguistics (Bres et al., 2016), which frames all discourse as interacting with other discourses such as the lower courts, parties and dissenting justices; with itself, through precedent; and with potential future discourses, such as members of the legal community affected by the decision.

The creation of a new coding scheme for SCOTUS opinions, with the goal of annotating finer-grained discourse structures and targeting the linguistic specificity of a corpus of American judicial opinions, presented an opportunity, among others, to remedy these issues in the current move analysis literature. The move analysis developed in this study is based on the current methodological standards which guarantee scientific rigor regarding the move analysis method (Kim et al., 2024) and which allow for reliable NLP processing of collaborative annotation (Fort, 2012). Following Moreno and Swales (2018), the aim was to construct abstract discourse structures (moves) from the identification of lower discourse units (steps).

Importantly, while our framework draws conceptually from prior work in rhetorical and discourse analysis, including Rhetorical Structure Theory (Mann & Thompson, 1988), we do not undertake a direct empirical benchmarking against RST-based annotation schemes or those used in existing legal NLP corpora. This choice is partly pragmatic, given the distinct goals of our annotation scheme —specifically, its alignment with genre-specific communicative functions rather than relational or hierarchical structure.

3. The annotation pre-campaign

To ensure the productivity and quality of the annotation campaign, we adopted a methodological approach inspired by Fort (2012). The entire process is illustrated in Appendix. We began with a preparatory phase involving four main participants: two campaign supervisors, also referred to as experts, with one being a native speaker of English (PhD holder) trained in linguistics, and the other being a native speaker of French trained in language teaching/learning (MA holder, C2 level in English); two "testers",

with one being a native speaker of German trained in linguistics (PhD holder, C2 level in English) and the other being a native speaker of Russian trained in language teaching/learning (PhD holder, C2 level in English). The pre-campaign phase included three main stages: first, a pilot stage consisting of data collection and the development of a coding scheme for steps in SCOTUS opinions (section 3.1); second, the creation of annotation guidelines and the annotation of a reference sample (section 3.2); and third, a verification of the annotation guide and a revision of the reference annotation (section 3.3).

3.1. Pilot phase

In this section, we report the process of compiling a set of representative samples of SCOTUS opinions. These texts are readily available and free from copyright restrictions, facilitating their use in creating educational resources. This accessibility not only simplifies the process of resource development but also ensures that learners have unrestricted access to important legal documents. In our study, we specifically concentrate on majority opinions because they represent the conclusive judgment and legal reasoning adopted by the majority of the Court.

Selection criteria of the documents include diversity of authorial perspectives, historical range of legal opinions, and topics. The aim of this dataset is to support machine learning and the evaluation of automatic recognition systems. The challenge was to define samples large enough to guarantee the representativeness of our criteria and, at the same time, not so large that the manual annotation task became unfeasible. In order to have opinions that were representative of the recent historical period and written in contemporary legal American English, the publication period was limited to 1945-2020. Based on a CourtListener[1] dump, this period represents 7,157 SCOTUS majority opinions written by 38 distinct justices.

3.1.1. Collecting representative samples

This substantial collection allows for observation of longitudinal changes and for an overall picture of judicial perspectives. The three sampling criteria (historical diversity, author diversity, and topical diversity) were originally identified in the effort to create a representative sample of the larger collection of opinions. Figure 1 illustrates a timeline showing the correlation between the authors (justices) of these opinions and the periods in which

they were active (midpoint). It highlights the considerable influence of different justices over the decades, and shows a consistent and balanced representation of authors over this long period.

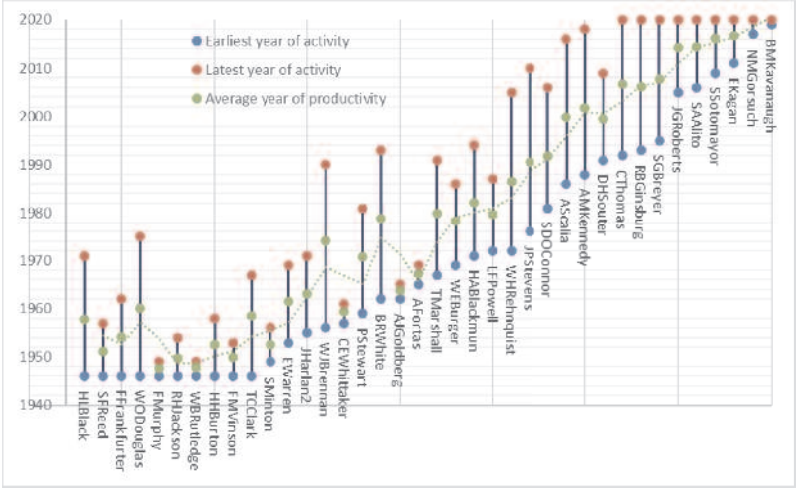


Figure 1. Timeline of judicial authorship for SCOTUS opinions (1945-2020).

The correlation between the "Author" and "Average Year of Productivity" criteria, that is, for each author, the weighing of each year of activity by the number of opinions written in that year, was calculated and deemed strong (Pearson coefficient = 0.88). It was therefore possible to use only the "Author" criterion, as using both could lead to overemphasizing the aspect of publication date in the sampling process. This also implies that, despite overlapping tenures, it is possible to distinctly evaluate each justice's contributions over time. As a result, ensuring equitable inclusion of the 38 justices who authored opinions during the chosen period ensures an even distribution across the timeline.

Our second criterion was to ensure thematic representativeness in terms of lexical diversity. Insofar as we did not initially know the nature of the steps present in the SCOTUS opinions, our intention was to compile situations where the legal language might differ and therefore possibly the nature of the steps taken. To systematically classify these opinions by thematic content so that they could be generalized, we used the K-Means (Ahmed et al., 2020) clustering algorithm using term frequency inverse document frequency (TF-

IDF) to group opinions into coherent thematic clusters based solely on their lexical content[2]. After preprocessing the corpus (including the removal of stop words and the conversion of all text to lowercase), multiple experiments were conducted to determine the optimal number of clusters. These experiments led us to establish that 18 clusters provided robust differentiation between documents, as illustrated in Figure 2.

Each cluster groups the opinions around a central thematic core, ensuring that all opinions within a cluster share thematic similarities.

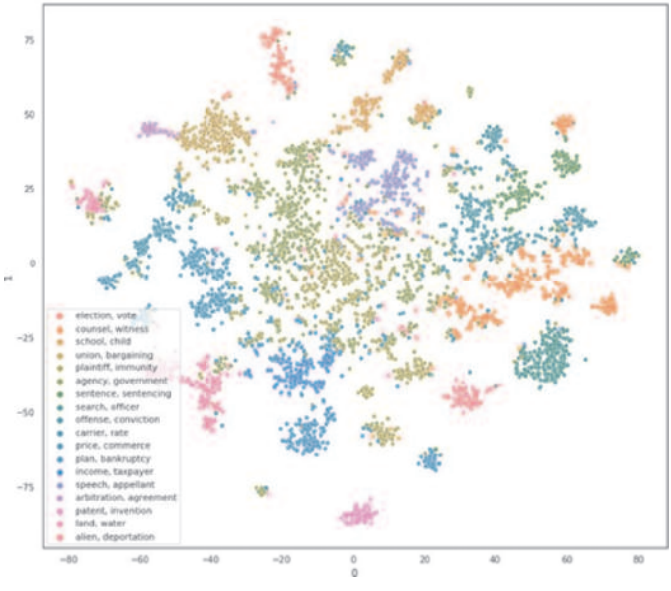


Figure 2. Cluster analysis of judicial topics in SCOTUS opinions.

Note. The cluster labels are based on the two most frequent words in each thematic cluster.

The selection of opinions covering the 18 themes for each author would have required, maximally, the manual annotation of 684 separate opinions (18 themes x 38 authors). To address this challenge more efficiently, we adopted a strategy that considers the dimensions of author and theme, while implementing a thresholding approach to select the most productive authors for each theme. For instance, some authors are more productive in their writing opinions about issues in criminal law, while others focus primarily on financial issues. After several experiments, we established that a threshold of four would allow for coverage of all the themes studied over the selected

period. In other words, if a justice wrote at least four opinions on a specific theme, this justice was deemed relevant for the theme, that is, their opinions could be selected to represent it. This strategy enabled us to extract samples of opinions representative of the themes and authors of the period.

Ultimately, we constructed a dataset comprising 10 samples, each containing 18 different opinions that adhered to the principles of balanced corpus construction. This method ensured a representative collection that adhered to feasibility constraints.

3.1.2. Development of the coding scheme

To annotate this dataset for moves and steps, the two experts sought to create a "workable" scheme. Another objective was obtaining a satisfactory number of occurrences for each label to train language models. The two experts first proposed labels based on the inductive analysis of seven SCOTUS majority opinions. They categorized the segments into functional-semantic units, as described in Cotos et al. (2017, pp. 94-95). The initial list of candidate steps contained 107 labels. Meetings between the two researchers allowed reducing the number of steps. Several strategies were used during this exploratory phase, such as merging labels with close rhetorical purpose or eliminating rarely used labels (see Supplementary Materials A for a complete description of the strategies).

At the end of this process, the two experts retained 35 step labels. Before having defined annotation guidelines, the experts tested the coding scheme to ensure that the labels could be consistently applied by distinct annotators on the collected representative samples. They separately annotated 18 majority opinions from the reference sample, resulting in 2,529 commonly annotated segments. They achieved a Cohen's kappa value of 0.67 in their annotation of these segments. The coefficient accounts for chance agreement, rather than the simple agreement rate and is therefore a more robust measure than percentages (Kanoksilapatham, 2007). Many move analysts use Landis and Koch (1977) benchmark to interpret Cohen's kappa values (Rau & Shih, 2021), according to which a 0.67 value amounts to a substantial agreement. It is lower than the reference value of 0.8 generally recommended by Artstein and Poesio (2008) to guarantee solid annotation quality. However, these authors warn against a universal threshold, stating that "[f]or some CL [computational linguistics] studies, particularly on discourse, useful corpora have been obtained while attaining reliability only

at the 0.7 level" (Artstein & Poesio, 2008, p. 591). Since this agreement value was obtained during the formative phase of a complex task— the analysis of the discourse structure of a newly studied genre— it aligned closely with the benchmark proposed by Artstein and Poesio (2008). Although a higher agreement would provide greater assurance for downstream NLP applications, the achieved score implied that the coding scheme used at this stage was suitable as a basis for subsequent annotation efforts.

3.2. Creation of the annotation guide and of an annotated reference sample

3.2.1. Defining and delimiting steps

Existing methodological frameworks for move analysis require annotating texts pertaining to a certain genre by identifying first moves (e.g., Swales, 1990), or steps (e.g., Moreno & Swales, 2018). We follow the latter approach but, in both approaches, the identification of these rhetorical divisions is primarily based on functional rather than formal criteria. Moreno and Swales (2018) argue that, from a syntactic point of view, a "step may be realized by a proposition, a complex of propositions or an even larger text fragment [...] and contain at least one verb, whether finite, non-finite or elliptical, or a nominalization easily convertible into a verbal phrase" (p. 49).

However, in the context of a coding scheme designed to be exploited by automatic language processing models, and with a view to accurately measuring inter-rater reliability (IRR) scores, the approach adopted for the task described in this paper was different. On the one hand, discourse criteria are difficult to associate with syntactic features. For example, many studies in SDRT use the clause as a minimal unit, but the existence of non-propositional syntactic units (e.g., adverbials) with discursive autonomy and a communicative function has also been highlighted, particularly if they are parenthetical (Muller et al., 2012). These considerations show the complexity of creating a segmentation paradigm that would cover all possible syntactic-discursive configurations. On the other hand, pilot experiments in the development stage (section 3.1.2) revealed divergent interpretations regarding step boundaries, even using the full definition provided by Moreno and Swales (2018). The segmentation approach adopted, therefore, was based primarily on the formal criterion of the sentence, as described in the next subsection of this paper. In some cases, however, the segmentation may occlude one or more secondary communicative functions appearing within a

sentence, and thus, this choice constitutes a divergence from the functional orientation of Swalesian move analysis.

3.2.2. Segmentation into elementary discourse units

The creation of the segmentation rules was guided by three principles: Segments in the corpus had to be uniform and ideally contain only one communicative function; each segment had to have a unique label to reduce complexity and avoid inconsistency; and the segmentation had to be automatically reproducible to ensure mass processing.

We chose, therefore, the syntactic sentence as the basic unit for segmentation and labelling. However, sentences containing several complete clauses can present multiple rhetorical functions. For this reason, specific rules were elaborated (see Supplementary Materials B) and a rule-based parser (to be released in open-access on the Lexhnology website) was developed to segment the opinions in the dataset. In total our dataset comprises 26,328 segments, i.e., a mean of 146 segments per opinion and 2,633 per sample.

3.2.3. Development of annotation guidelines

The two experts designed the annotation guide to cover the selection of a step label. The guide was based on the inferred communicative function of each segment. The process of selecting a step label was divided into several successive stages to simplify the annotation process (see Figure 3). Each step label identified and verified during the creation of the annotation guide was divided into a discursive category, a rhetorical function, and possible attributes (type, target and author). The categories reflect the current organization of SCOTUS opinions. These generally include a part introducing the case's background and how it arose (*Setting the scene*), a part where the Court analyzes the issue and justifies its reasoning (*Analysis*), and a concluding part (*Resolution*). A fourth category, *Sources of Authority* includes any textual segments referring to sources of authority throughout the texts, while the last category (*Announcing function*) only includes one rhetorical function with text-organizing purposes.

The rhetorical function identifies the communicative purpose of the segment to be annotated. In addition, attributes are used to complement the rhetorical function, providing information about the type (i.e., the elements being described or recalled, or the type of source of authority being

mentioned), author (i.e., the originator of the argument being recalled), and target (i.e., distinguishing between information related to the adjudicated case or other former cases).

An American lawyer also verified the coding scheme and guide during the preparation for the annotation campaign. She worked with the two experts to ensure that the annotation respected principles of American law. The final version of the coding scheme, after minor revisions during the creation and test of the annotation guide, is presented in Figure 3. Table 1 flattens this coding scheme and reports the actual number of annotated segments. Information about the content of the annotation guide is provided in Supplementary Materials C.

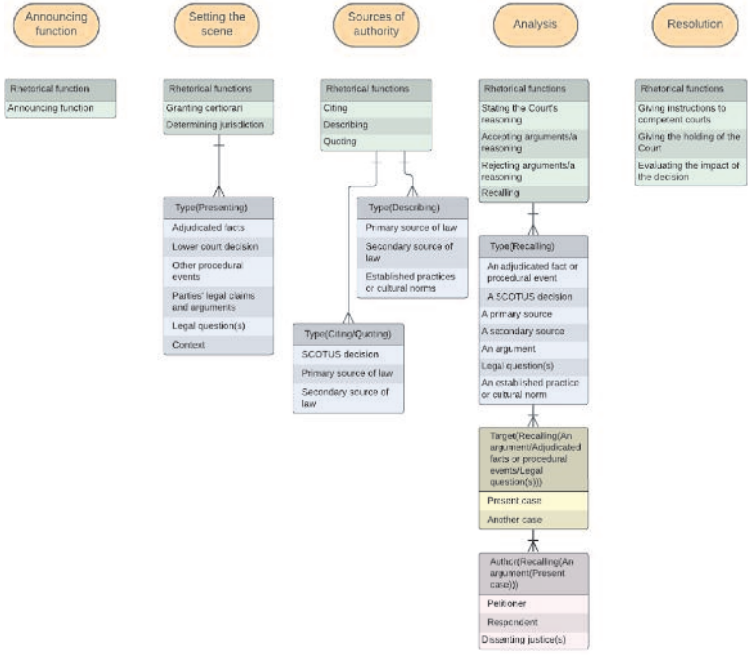


Figure 3. Final coding scheme for steps in SCOTUS opinions.

Note. The final coding scheme is composed of 5 categories (ovals with orange background), 13 rhetorical functions (green rectangles) and 24 attributes (types in blue rectangles, target in the yellow rectangle, and author in the purple rectangle). The scheme reads from top to bottom, which reflects the annotation process in the annotation guide and in Label Studio: A step label is constructed by first choosing a category, then a rhetorical function, then if required, by combining attributes to complete the discursive information provided by the rhetorical function.

Category		Rhetorical Function		Type		Target		Author	
Announcing function	344	Announcing function	344						
Setting the scene	5123	Granting certiorari	182						
		Presenting	4941	Adjudicated facts	2283				
				Lower court decision	1192				
				Context	467				
				Other procedural events	412				
				Parties' legal claims and arguments	363				
				Legal question(s)	224				
Sources of authority	8041	Citing	6442	SCOTUS decision	2764				
				Primary source of law	2203				
				Secondary source of law	1474				
		Describing	955	Primary source of law	771				
				Secondary source of law	159				
				Established practices or cultural norms	25				
		Quoting	644	SCOTUS decision	235				
				Primary source of law	241				
				Secondary source of law	168				
Analysis	11910	Stating the Court's reasoning	3198						
		Rejecting arguments/a reasoning	490						
		Accepting arguments/a reasoning	103						
		Recalling	8119	A SCOTUS opinion	2160				
				A primary source	1781				
				A secondary source	359				
				An established practice or cultural norm	1199				
				An adjudicated fact or procedural event	1447	Present case	1152		
						Another case	295		
				Legal question(s)	182	Present case	147		
						Another case	35		
				An argument	991	Present case	967	Petitioner	413
								Respondent	513
								Dissenting justice(s)	22
						Another case	24		
Resolution	910	Giving the holding of the Court	760						
		Giving instructions to competent courts	105						
		Evaluating the impact of the decision	45						
Total	26328								

Table 1. Flattened coding scheme with counts of occurrences for each element.

3.2.4. Reference annotation

The experts proceeded to annotate the reference sample following the annotation guidelines to create a gold standard for the annotation. The gold standard is considered prioritized over the performance of other human annotators or computer models.

The web-based annotation software Label Studio (Tkachenko et al., 2020) was chosen because of its open-source features. It offers cloud and collaborative capabilities in terms of project management, quality control

and report. Importantly, it also allows for the annotation of HTML documents. Conserving the visual structure, which allows for easier navigation through the documents, was important for the annotation process because of the documents' length and complexity. Finally, Label Studio also allows for management of multiple annotators as well as the assignment of different roles. A template proposed by the software was chosen for annotation but required several versions to represent the complex coding scheme in Figure 3. This template may be found in Supplementary Materials D.

At the conclusion of this process, the IRR was calculated for the two experts using Cohen's kappa value. The two experts demonstrated a good level of coherence, with a Cohen's kappa value of 0.67. Disagreements were resolved and one unique label was selected for each annotated segment.

3.3. Verification of the annotation guide

In order to ensure that the information in the guide could be understood and applied by the future annotators of the corpus, two testers were integrated. The testers were one linguist and one specialist of teaching and learning foreign languages. An iterative protocol was developed to identify and improve sections of the annotation guide that were unclear (see Supplementary Materials E). A major change was inspired by the coding scheme proposed by Teufel et al. (1999) for moves in research articles. Following these authors' approach, discrete criteria were added to the guide to distinguish between argumentative segments referencing past own work (OWN in Teufel's scheme), theoretical context (BACKGROUND in Teufel's scheme), or offering authoritative support for the author's propositions (BASIS in Teufel's scheme).

For example, the rhetorical function *Recalling* was used for the reference to authoritative sources (i.e., BASIS) providing support to original arguments from the Court (*Stating the Court's reasoning*, comparable to OWN in Teufel's scheme).

The guide was revised after each of the six opinions annotated by the two testers. At the end of the process no major confusion remained. The entire reference sample was then reannotated by the two experts to ensure consistency of the gold standard used for the annotation phase described in Section 4.

4. The annotation campaign

Two annotators with a background in both ELP and law were selected from an undergraduate program. During their training, they annotated part of the reference sample then 9 other samples in the dataset. We describe the training and annotation phases in the following sections. We also briefly present the resource itself.

4.1. Annotation

After recruitment, the two annotators participated in two weeks of theoretical and practical training to master the annotation process. The training materials are described in Supplementary Materials F. The large-scale annotation phase lasted two months. The two experts coordinated the campaign and assigned majority opinions of the representative samples in the SCOTUS corpus to the two annotators in Label Studio as tasks. Task assignment between the two annotators was random. Each received 9 tasks per sample. The annotators were in the same room, and were allowed to discuss and resolve problems of text comprehension. In addition, the two experts held daily meetings with them to discuss any remaining difficulties, as well as to review parts of the annotation that seemed questionable at first inspection. Throughout the campaign, the IRR between the annotators was tested to ensure that they were still achieving a good level of coherence with each other and also with the experts. The first two tests were announced. The annotators were asked to annotate two opinions from the reference sample that were new to them. The last two tests involved opinions drawn from other samples, in which both annotators were assigned to the same task without their knowledge. IRR rates revealed no major differences between the two tests, and levels of consistency were considered good: the Fleiss Kappa values reached 0.72 among the two experts and the two annotators on the reference sample.

4.2. Preliminary exploitation of the annotated corpus

At the end of the annotation process, the 10 samples which composed the SCOTUS dataset were entirely annotated. This corresponds to a total of 180 texts (18 opinions per sample). This accounted for 26,328 annotated segments as shown in Table 1, with a significant representation for the categories *Analysis* and *Sources of Authority*, with 11,910 and 8,041 annotations, respectively.

In Figure 4 and Figure 5, we illustrate the distribution at both the category and rhetorical function levels. We observe an imbalance in terms of the frequency with which one type of category appears, indicating a major presence for some categories over others. The same observation applies to rhetorical functions. However, the standard deviation between the groups and the entire corpus, is low, for example, ±1.93% for *Setting the scene*. This confirms the effectiveness of our dimension choices in characterizing representativeness and ensuring diversity in the annotated opinions, as discussed in Section 3.

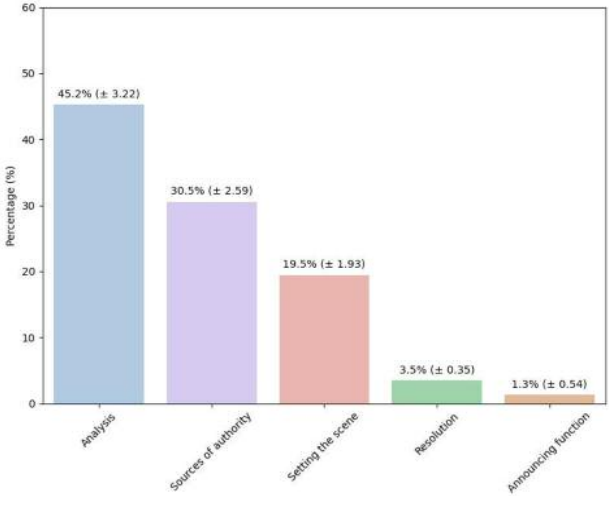


Figure 4. Distribution of categories across the corpus.
Note. The (± X) above bars indicates the average standard deviation within the distribution of each label (category, rhetorical function) relative to the total.

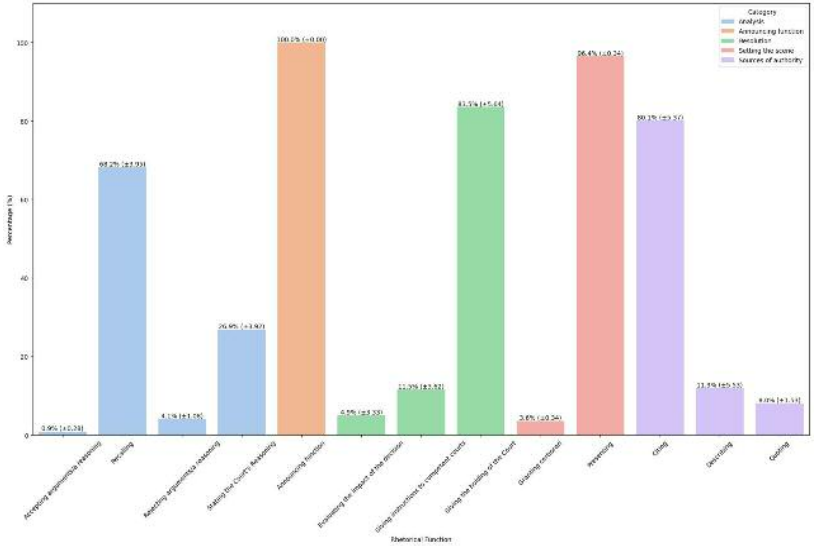


Figure 5. Distribution of rhetorical functions across categories in the corpus.
Note. The (± X) above bars indicates the average standard deviation within the distribution of each label (category, rhetorical function) relative to the total.

5. Discussion

The move analysis of U.S. legal opinions described in this article was initially motivated by pedagogical goals. However, the challenges encountered during the implementation of move analysis provided an opportunity to contribute methodologically to the field of English for Specific Purposes, particularly English for Legal Purposes, regarding the following research question: How can move analysis methodology be adapted for machine learning and legal discourse?

5.1. Specific constraints of NLP compatible move analysis

While the emergence of large language models (LLMs) has reignited interest in the automated detection of moves in specialized genres, the application of supervised machine learning techniques in this area remains limited. Such models require extensive training on manually annotated datasets, which is both time-intensive and costly.

Before annotating a corpus, it is crucial to develop an annotation framework that yields results compatible with NLP models. While Moreno and Swales

(2018) emphasize reducing the number of labels to make annotation manageable for human annotators, automated approaches require an additional consideration: each label must appear frequently enough to enable the model to learn the linguistic features associated with the corresponding rhetorical unit. This challenge is highlighted by Anthony and Lashkia (2003), who point out that the structural realities of a genre —where some communicative functions are underrepresented compared to others— can hinder NLP models from effectively identifying the less common functions. To mitigate this issue, it is essential to design an annotation framework in which more granular and less frequent rhetorical functions are embedded within broader rhetorical categories, as proposed here.

Another key adaptation involves the segmentation of rhetorical units within texts. Traditional approaches segment texts into distinct communicative units, assigning a rhetorical function to each based on the annotation framework. This preliminary task relies heavily on the cognitive judgment of annotators, which poses challenges for inter-annotator agreement, as noted by Cotos et al. (2017). Furthermore, automated systems often struggle with segmentation. For instance, Dayrell et al. (2012) were unable to develop a system that could automatically segment texts into move-compliant units. Consequently, we opted for an automated segmentation approach based on syntactic rather than functional criteria, segmenting texts at the sentence level.

Regarding the potential multifunctionality of rhetorical units, we chose to assign a single label per sentence, also following Moreno and Swales (2018), even though it is recognized that segments can carry more than one rhetorical function (Cotos et al., 2017). Recent studies in move analysis suggest that when annotators are allowed to assign multiple rhetorical functions to a single unit, they tend to agree on the primary function but often disagree on secondary ones (Kim et al., 2024, p. 12). This would create disagreement in the datasets and thus may lead models to learn incorrect associations. Models trained to apply multiple labels can also struggle to identify the dominant rhetorical function of a unit (Dayrell et al., 2012).

5.2. Specific constraints of performing a move analysis on legal discourse

The development of this framework also required addressing the specificities of the genre studied here: SCOTUS majority opinions. These

texts are lengthy and complex, with their organizational features only vaguely described in both genre analysis research (Han et al., 2018) and professional literature (Lavissière & Bonnard, 2024). These characteristics present an obstacle to understanding how SCOTUS justices construct their argumentation, highlighting the need for a novel framework capable of capturing a higher level of rhetorical complexity.

One of the most immediate challenges in applying existing move/step frameworks to SCOTUS majority opinions is that a single move, such as *Arguing the Case*, could account for up to 90% of the text and extend over 15,000 words in longer opinions. Moves are defined as abstract units serving a rhetorical purpose (Swales, 1990) and are considered "flexible in terms of their linguistic realization" (pp. 228-229). However, Moreno and Swales (2018) observe that moves are typically identified at levels ranging from clauses to paragraphs. Swales' definition theoretically allows a move to span multiple paragraphs. This broad scope is problematic for one of the primary pedagogical objectives of move analysis —providing a schematic representation to illustrate the practices of specific discourse communities.

This raises questions about the applicability of move analysis for describing the structure of certain specialized genres, particularly those that are lengthy and lack standardized discourse segmentation at a finer level. Within the broad rhetorical purposes identified in judicial discourse (Mazzi, 2007), it is difficult to isolate sufficiently recurrent discursive patterns at the *step* level to construct a meaningful schematic representation of the genre (see Lavissière & Bonnard, 2024).

Some rhetorical elements of judicial opinions are well-documented in professional literature. For instance, van Geel (2009) notes that SCOTUS justices rely on a wide array of sources to support their reasoning, including the US Constitution, societal practices and traditions, and SCOTUS precedents. Each of these sources is associated with distinct rhetorical strategies in judicial argumentation. For example, when citing precedent, justices may employ strategies such as *Drawing a factual distinction*, *Narrowly interpreting the ratio decidendi*, *Reducing an announced principle to obiter dictum*, or *Rejecting a fact as material* (van Geel, 2009). While all these strategies could theoretically be categorized as steps under the move *Arguing the Case*, the sheer number of potential steps would be overwhelming considering all possible sources and related rhetorical strategies that SCOTUS justices employ in their reasoning. Additionally, many of these plausible steps would

occur too infrequently in a manually annotated corpus to be analytically useful in an NLP context.

To address these challenges, we developed an innovative discourse annotation framework that departs somewhat from traditional move analysis principles by integrating multiple layers of rhetorical information. The framework consists of:

1. Categories: These correspond roughly to the major sections identified in genre analysis and professional legal literature, but we do not consider them moves due to their extensive scope.
2. Rhetorical Functions and Types of materials: These refine the nature of more specific communicative units within the broader categories. They could be considered steps under Swales' definition, but some of the rhetorical functions in our scheme are completed by essential information to characterize legal discourse, particularly the sources (materials) justices rely on. For example, in the case of the rhetorical function *Recalling*, the annotation scheme specifies various types of legal authority, such as an *established practice or cultural norm*, a *SCOTUS decision*, or a *primary source*.
3. Elements to identify the interplay of voices: These provide additional characterization of certain rhetorical functions, accounting for the dialogic nature of judicial discourse, where multiple voices interact within the text. For instance, for the rhetorical function *Recalling* in the category *Analysis*, the annotation scheme includes an attributive element identifying the author of the recalled argument. This approach echoes Mazzi's (2007) model for ECJ decisions.

By integrating these additional layers, our framework offers a more nuanced approach to judicial discourse analysis, aligning with both theoretical considerations and practical needs in legal education and research.

6. Conclusion and future work

Our objective in this paper is to contribute to methodological literature about move analysis in applied linguistics and ESP. We provide the details of the development of a resource for legal discourse, specifically that of

SCOTUS. We include descriptions of the framework and processes used to develop the resource, the intermediate products of the development process, such as the coding scheme and annotation guide. The measures related to inter-rater reliability indicate that good consistency was achieved in a difficult discourse analysis annotation task. To our knowledge, this is the only resource in the legal field annotated using the Swalesian approach at this granularity and in this quantity.

While the framework is well suited to SCOTUS majority opinions, its generalizability to other legal genres remains limited. Indeed, it was specifically developed for a type of legal discourse in which the Court occupies the central position in a dynamic interplay of judicial voices. Grounded in genre analysis theory, it thus takes into accounts the social, institutional and communicational constraints that shape SCOTUS judicial opinions. Accordingly, applying the framework to other legal contexts (e.g., lower court decisions or rulings from other legal traditions) would require adaptation to reflect different rhetorical conventions and institutional roles. Furthermore, it is not intended for non-argumentative legal genres such as legislation or contracts. In future work, we aim to explore the adaptability of the coding scheme in various judicial contexts and refine it for broader use in legal discourse analysis.

We also plan to use this resource in several ways: exploring the presence of moves via NLP techniques, annotating a larger corpus of SCOTUS opinions, and evaluating the impact of highlighting discourse units on reading comprehension.

Acknowledgements

We would like to express our gratitude to the editor and anonymous reviewers for their valuable comments on this paper. We are also grateful to Léo Lisana and Déborah Boudzoumou for their much appreciated assistance with corpus annotation. This research was funded by the French National Research Agency, grant number ANR-22-CE38-0004.

Article history:
Received 26 March 2025
Received in revised form 15 June 2025
Accepted 15 June 2025

References

Ahmed, M., Seraj, R., & Islam, S. M. S. (2020). The k-means algorithm: A comprehensive survey and performance evaluation. *Electronics*, *9*(8), 1295. https://doi.org/10.3390/electronics9081295

Anthony, L., & Lashkia, G. V. (2003). Mover: A machine learning tool to assist in the reading and writing of technical papers. *IEEE Transactions on Professional Communication*, *46*(3), 185-193. https://doi.org/10.1109/tpc.2003.816789

Artstein, R., & Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational linguistics*, *34*(4), 555-596. https://doi.org/10.1162/coli.07-034-r2

Asher, N., & Lascarides, A. (2003). *Logics of conversation*. Cambridge University Press.

Bhatia, V. K. (1993). *Analysing genre: Language use in professional settings*. Routledge.

Biber, D., Connor, U., & Upton, T. A. (2007). *Discourse on the Move: Using corpus analysis to describe discourse structure*. John Benjamins. https://doi.org/10.1075/scl.28

Boulton, A., Kalyaniwala, C., & Lavissière, M. C. (2025). Teaching reading for case law in English for Legal Purposes. *ASp. La revue du GERAS*, *87*, 71-94. https://doi.org/10.4000/13pod

Breeze, R. (2009). Issues of persuasion in academic law abstracts. *Revista Alicantina de Estudios Ingleses*, *22*, 11-26. https://doi.org/10.14198/raei.2009.22.02

Bres, J., Nowakowska, A., & Sarale, J.-M. (2016). Anticipative interlocutive dialogism: Sequential patterns and linguistic markers in French. *Journal of Pragmatics*, *96*, 80-95. https://doi.org/10.1016/j.pragma.2016.02.007

Casal, J. E., & Kessler, M. (2023). Rhetorical move-step analysis. In M. Kessler & C. Polio (Eds.), *Conducting genre-based research in applied linguistics* (pp. 82-104). Routledge. https://doi.org/10.4324/9781003300847-7

Cotos, E., Huffman, S., & Link, S. (2015). Furthering and applying move/step constructs: Technology-driven marshalling of Swalesian genre theory for EAP pedagogy. *Journal of English for Academic Purposes*, *19*, 52-72. https://doi.org/10.1016/j.jeap.2015.05.004

Cotos, E., Huffman, S., & Link, S. (2017). A move/step model for methods sections: Demonstrating Rigour and Credibility. *English for Specific Purposes*, *46*, 90-106. https://doi.org/10.1016/j.esp.2017.01.001

Dayrell, C., Candido Jr, A., Lima, G., Machado Jr, D., Copestake, A. A., Feltrim, V. D., Tagnin, S. E., & Aluísio, S. M. (2012). Rhetorical move detection in English abstracts: Multi-label sentence classifiers and their annotated corpora. In *Proceedings of the Eighth International Conference on language resources and evaluation (LREC'12)* (pp. 1604-1609). ELRA.

Dudley-Evans, T. (2002). Genre analysis: An approach to text analysis for ESP. In M. Coulthard (Ed.), *Advances in written text analysis* (pp. 233-242). Routledge.

Fort, K. (2012). *Les ressources annotées, un enjeu pour l'analyse de contenu : Vers une méthodologie de l'annotation manuelle de corpus*. Doctoral dissertation. Université Paris 13.

Gozdz-Roszkowski, S. (2020). Move analysis of legal justifications in Constitutional Tribunal judgments in Poland: What they share and what they do not. *International Journal for the Semiotics of Law – Revue Internationale de Sémiotique Juridique*, *33*(3), 581-600. https://doi.org/10.1007/s11196-020-09700-1

Groom, N., & Grieve, J. (2019). The evolution of a legal genre. In T. Fanego & P. Rodriguez-Puente (Eds.), *Corpus-based research on variation in English legal discourse* (pp. 201-234). John Benjamins. https://doi.org/10.1075/scl.91.09gro

Han, Z. (2011). The discursive construction of civil judgments in Mainland China. *Discourse & Society*, *22*(6), 743-765. https://doi.org/10.1177/0957926511419924

Han, Z., Bhatia, V. K., & Ge, Y. (2018). The structural format and rhetorical variation of writing Chinese judicial opinions: A genre analytical approach. *Pragmatics*, *28*(4), 463-488. https://doi.org/10.1075/prag.17013.ge

Joty, S., Carenini, G., Ng, R., & Murray, G. (2019). Discourse analysis and its applications. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts* (pp. 12-17). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-4003

Kalamkar, P., Tiwari, A., Agarwal, A., Karn, S., Gupta, S., Raghavan, V., & Modi, A. (2022). *Corpus for automatic structuring of legal documents*. http://arxiv.org/abs/2201.13125

Kanoksilapatham, B. (2007). Rhetorical moves in biochemistry research articles. In D. Biber, U. Connor & T. A. Upton (Eds.), *Discourse on the Move: Using corpus analysis to describe discourse*

structure (pp. 73-119). John Benjamins. https://doi.org/10.1075/scl.28.06kan

Kim, M., Qiu, X., & Wang, Y. (2024). Interrater agreement in genre analysis: A methodological review and a comparison of three measures. *Research Methods in Applied Linguistics*, *3*(1), 100097. https://doi.org/10.1016/j.rmal.2024.100097

Kirby-Légier, C. (2005). Understanding the judicial discourse of the current United States Supreme Court. In R. Greenstein (Ed.), *La langue, le discours et la culture en anglais du droit* (pp. 87-110). Publications de la Sorbonne.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 159-174.

Larsson, T., Plonsky, L., Sterling, S., Kytö, M., Yaw, K., & Wood, M. (2023). On the frequency, prevalence, and perceived severity of questionable research practices. *Research Methods in Applied Linguistics*, *2*(3), 100064. https://doi.org/10.1016/j.rmal.2023.100064

Lavissière, M. C., & Bonnard, W. (2024). Who's really got the right moves? Analyzing recommendations for writing American Judicial Opinions. *Languages*, *9*(4), 119. https://doi.org/10.3390/languages9040119

Lavissière, M. C., Boulton, A., & Bonnard, W. (2024). Integrating a Swalesean move analysis in reading legal texts. *Études en Didactique des Langues*, *42,* 31-53.

Le, T. N. P., & Pham, M. M. (2020). Genre practices in mechanical engineering academic articles. *Ibérica, Journal of the European Association of Languages for Specific Purposes, 39*, 243-266. https://doi.org/10.17398/2340-2784.39.243

Lexhnology (n.d.). *Publications*. https://lexhnology.hypotheses.org/publications

Mann, W. C., & Thompson, S. A. (1988). Rhetorical Structure Theory: Toward a functional theory of text organization. *Text – Interdisciplinary Journal for the Study of Discourse*, *8*(3), 243-281. https://doi.org/10.1515/text.1.1988.8.3.243

Mazzi, D. (2007). The construction of argumentation in judicial texts: Combining a genre and a corpus perspective. *Argumentation*, *21*(1), 21-38. https://doi.org/10.1007/s10503-007-9020-8

Moreno, A. I., & Swales, J. M. (2018). Strengthening move analysis methodology towards bridging the function-form gap. *English for Specific Purposes*, *50*, 40-63. https://doi.org/10.1016/j.esp.2017.11.006

Muller, P., Vergez-Couret, M., Prévot, L., Asher, N., Farah, B., Bras, M., Le Draoulec, A., & Vieu, L. (2012). *Manuel d'annotation en relations de discours du projet Annodis*. http://www.irit.fr/~Philippe.Muller/perso_utf8_bib.html

Rau, G., & Shih, Y.-S. (2021). Evaluation of Cohen's kappa and other measures of inter-rater agreement for genre analysis and other nominal data. *Journal of English for Academic Purposes*, *53*, 101026. https://doi.org/10.1016/j.jeap.2021.101026

Salager-Meyer, F. (1992). A text-type and move analysis study of verb tense and modality distribution in medical English abstracts. *English for Specific Purposes*, *11*(2), 93-113. https://doi.org/10.1016/S0889-4906(05)80002-X

Soler Monreal, C. (2016). A move-step analysis of the concluding chapters in computer science PhD theses. *Ibérica, Journal of the European Association of Languages for Specific Purposes*, *32*, 105-132. https://revistaiberica.org/index.php/iberica/article/view/175

Swales, J. M. (1990). *Genre analysis*. Cambridge University Press.

Swales, J. M. (2004). *Research genres: Explorations and applications*. Cambridge University Press.

Tessuto, G. (2021). Making sense of web-based European Court of Justice institutional press releases: Context, structure and replicable genres. *Ibérica, Journal of the European Association of Languages for Specific Purposes, 42*, 219-244. https://doi.org/10.17398/2340-2784.42.219

Teufel, S., Carletta, J., & Moens, M. (1999). An annotation scheme for discourse-level argumentation in research articles. In *Proceedings of the Ninth Conference on European Chapter of the Association for Computational Linguistics* (pp. 110-117). https://doi.org/10.3115/977035.977051

Tkachenko, M., Malyuk, M., Holmanyuk, A., & Liubimov, N. (2020). *Label Studio: Data labeling software*. https://labelstud.io

van Geel, T. R. (2009). *Understanding Supreme Court Opinions* (6th Edition). Routledge.

Yaich, M., & Hernandez, N. (2025). *Improving accessibility of SCOTUS opinions: A benchmark study and a new dataset for generic heading prediction and specific heading generation*. Association for Computational Linguistics.

Yang, R., Xu, L., & Swales, J. M. (2023). Tracing the development of *English for Specific Purposes* over four decades (1980-2019): A bibliometric analysis. *English for Specific Purposes*, *71*, 149-

160. https://doi.org/10.1016/j.esp.2023.04.004

Yu, D., Bondi, M., & Hyland, K. (2024). Can GPT-4 learn to analyse moves in research article abstracts? *Applied Linguistics*, amae071. https://doi.org/10.1093/applin/amae071

Warren Bonnard is a PhD student in applied linguistics at Université de Lorraine. His research interests include genre analysis, corpus linguistics, register variation and English for Specific Purposes.

Mary Catherine Lavissière, PhD, is Associate Professor in applied linguistics at Nantes Université. She holds degrees in English and Spanish linguistics from Sorbonne University. She publishes on language for specific purposes, notably via the prisms of genre, macrodivision, intertextuality, morphosyntax and diachrony. She has multiple publications in international journals on legal genres, the modernization of legal language, maritime discourse and genres, and textometry applied to management sciences. She is the principal investigator of Lexhnology ANR-22-CE38-0004, which aims to identify the rhetorical structure of case law through artificial intelligence methods for teaching and learning purposes. She co-coordinates Theme 2 Research Team ("National and transnational, legal, cultural and socio-economic environments") at her research institute, the Centre de Recherche sur les Identités, les Nations et l'Interculturalité (CRINI).

Anas Belfathi is a PhD Student at Nantes Université. His research questions concern the capture of rhetorical information from texts and its representation in language models. His application framework is the task of labelling sentence sequences to support applications in reading assistance or automatic summarisation.

Nicolas Hernandez is Associate Professor at Nantes Université. His research focuses on NLP and Text Mining, with a particular interest in applications to help humans access and understand textual content in specialized domains (e.g., legal, education). He is particularly interested in deep neural architectures and learning techniques for modeling and analyzing discourse structure.

Christine Jacquin is Associate Professor at Nantes Université. Her research expertise lies in NLP, with a particular interest in the development of annotated resources that are used for both training and evaluation of deep models. Her fields of application are related to health and education.

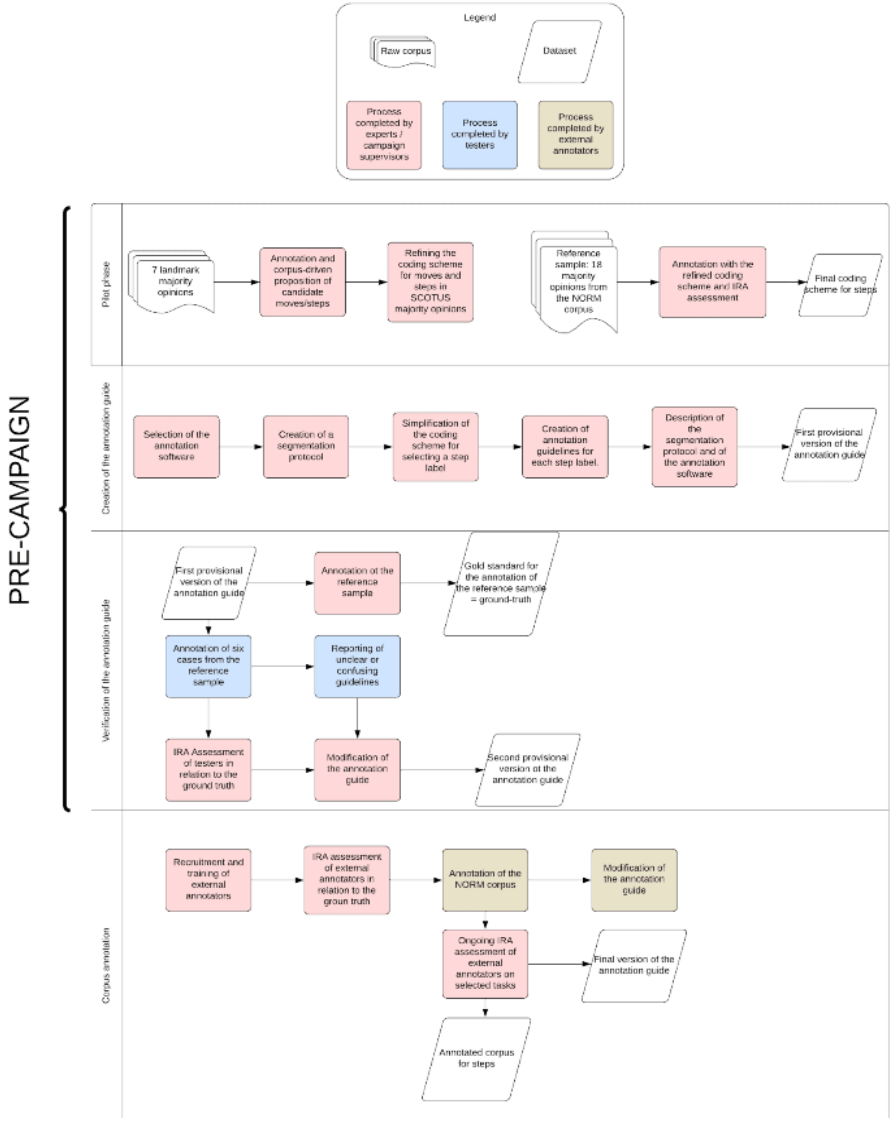
Laura Monceaux Cachard is Associate Professor at Nantes Université. Her current research lies within the field of NLP, particularly corpus annotation, discourse analysis, and language modeling, with a current emphasis on applications in the legal domain.

NOTES

[1] https://www.courtlistener.com

[2] This part of the study was carried out using a ready-to-use dataset available on Kaggle (https://www.kaggle.com/datasets/gqfiddler/scotus-opinions) which originally came from CourtListener over the same period.

Appendix: The annotation process



Supplementary Materials

A. Strategies to reduce labels

- Merging labels: Labels with similar or closely related communicative functions were grouped and merged into broader categories, especially when the verb was identical. For example, 'Accepting the petitioner's arguments', 'Accepting the respondent's arguments', and 'Agreeing with a lower court' were consolidated into 'Accepting arguments/a reasoning'.
- Eliminating rarely used labels: Labels that were rarely or never used in the annotation of majority opinions were discarded. Examples include 'Asserting the legitimacy of the decision' and 'Describing the type of case', which annotators eliminated in favor of labels with broader communicative functions.
- Removing labels which fell outside of the move analysis: Labels which described discursive relations between different parts of the texts, akin to the rhetorical relations in RST, were eliminated. Although they fulfill a discursive function, they are not formulated in terms of the main communicative goal of the genre. Labels such as 'Making a concession' and 'Introducing a consequence' were therefore abandoned as they would have required a double annotation of each segment.

B. Extra segmentation rules

- Paragraph breaks and punctuation marks like ":", or ";", unless within parentheses, acted as segment dividers.
- Segments with distinct communicative functions in separate independent clauses were divided if marked by "and," "or," or "but."
- Without these conjunctions, the main clause's function determined the label.
- When a sentence included both the fixed expression indicating that the Court agreed to examine the case as well as the statement of the case's legal issue (e.g., "We granted certiorari // to determine whether [. . .]"), these parts were divided into two segments for labelling.

C. Information about the annotation guide

The guide includes the following pieces of information for each rhetorical function:

- A detailed description of the contexts of use
- A list of attributes to be added and their explanations
- A set of linguistic and discursive clues to identify segments relating to the rhetorical function
- Examples in context
- A set of rules and decision and control procedures to be applied in case of doubt as to the label to be applied

The guide also explains how to use Label Studio (*Label Studio*, 2024), the annotation software chosen for the annotation campaign.

D. Template (with the labels in French) for the annotation of discourse units in the tagging software (LabelStudio)

<View style="display: flex;">

```
<View style="flex: 30%;">
  <Labels name="label" toName="text">
   <Label value="Mise en scène" background="blue"/>
   <Label value="Sources d'autorité" background="red"/>
   <Label value="Analyse" background="green"/>
   <Label value="Résolution" background="pink"/>
   <Label value="Fonction d'annonce" background="orange"/>
   <Label value="Sentence" background="grey"/>
  </Labels>
      <!— la catégorie Mise en scène —>
    <View visibleWhen="region-selected" whenLabelValue="Mise en scène">
        <Choices name="ch-1" showInline="true" toName="text" perRegion="true" choice="single" whenLabelValue="Mise en scène" required="true" requiredMessage="choix fonction rhétorique obligatoire">
            <Header value="Fonction rhétorique ?" underline="true"/>
            <Choice value="Accepter de revoir l'affaire"/>
            <Choice value="Exposer"/>
 </Choices>
        <Choices name="ch-1-1" toName="text" visibleWhen="choice-selected" whenTagName="ch-1" perRegion="true" whenChoiceValue="Exposer" choice="single">
            <Header value="Type ?" underline="true"/>
            <Choice value="Faits jugés"/>
            <Choice value="Décision d'une cour inférieure"/>
            <Choice value="Autres éléments de procédure"/>
            <Choice value="Arguments et prétentions des parties"/>
      <Choice value="Problème juridique"/>
      <Choice value="Contexte large"/>
  </Choices>

</View>
      <!— la catégorie Sources d'autorité —>
<View visibleWhen="region-selected" whenLabelValue="Sources d'autorité">
      <Choices name="ch-2" showInline="true" toName="text" perRegion="true" choice="single" whenLabelValue="Sources d'autorité" required="true" requiredMessage="choix fonction rhétorique obligatoire">
            <Header value="Fonction rhétorique ?" underline="true"/>
            <Choice value="Mentionner"/>
```

```
            <Choice value="Décrire le contenu"/>
            <Choice value="Citer un extrait"/>
  </Choices>
      <Choices name="ch-2-1" toName="text" visibleWhen="choice-selected" whenTagName="ch-2" whenChoiceValue="Mentionner" perRegion="true" choice="single">
            <Header value="Type ?" underline="true"/>
            <Choice value="Décision de la Cour Suprême"/>
            <Choice value="Source primaire de droit"/>
            <Choice value="Source secondaire de droit"/>
 </Choices>
      <Choices name="ch-2-2" toName="text" visibleWhen="choice-selected" whenTagName="ch-2" whenChoiceValue="Citer un extrait" perRegion="true" choice="single">
            <Header value="Type ?" underline="true"/>
            <Choice value="Décision de la Cour Suprême"/>
            <Choice value="Source primaire de droit"/>
            <Choice value="Source secondaire de droit"/>
  </Choices>
      <Choices name="ch-2-3" toName="text" visibleWhen="choice-selected" whenTagName="ch-2" whenChoiceValue="Décrire le contenu" perRegion="true" choice="single">
            <Header value="Type ?" underline="true"/>
            <Choice value="Source primaire de droit"/>
            <Choice value="Source secondaire de droit"/>
            <Choice value="Pratiques établies ou normes culturelles"/>
  </Choices>
</View>

  <!— la catégorie Analyse —>
  <View visibleWhen="region-selected" whenLabelValue="Analyse">
      <Choices name="ch-3" showInline="true" toName="text" perRegion="true" choice="single" whenLabelValue="Analyse" required="true" requiredMessage="choix fonction rhétorique obligatoire">
            <Header value="Fonction rhétorique ?" underline="true"/>
            <Choice value="Présenter le raisonnement de la Cour"/>
            <Choice value="Approuver un argument/un raisonnement"/>
            <Choice value="Rejeter un argument/un raisonnement"/>
           <Choice value="Rappeler"/>
```

```
 </Choices>

        <Choices name="ch-3-1" toName="text" visibleWhen="choice-selected" whenTagName="ch-3" whenChoiceValue="Rappeler" perRegion="true" choice="single">

            <Header value="Type ?" underline="true"/>

            <Choice value="Un fait jugé ou un élément de procédure"/>

            <Choice value="Une décision de la Cour"/>

            <Choice value="Une source primaire"/>

            <Choice value="Une source secondaire"/>

            <Choice value="Une pratique établie ou norme culturelle"/>

            <Choice value="Un argument"/>

             <Choice value="Problème juridique"/>


</Choices>

<Choices name="ch-3-1-1" toName="text" visibleWhen="choice-selected" whenChoiceValue="Un argument" whenTagName="ch-3-1" choice="single" perRegion="true">

    <Header value="Cible ?" underline="true"/>

    <Choice value="Affaire jugée"/>

    <Choice value="Autre affaire"/>

 </Choices>

   <Choices name="ch-3-1-2" toName="text" visibleWhen="choice-selected" whenChoiceValue="Un fait jugé ou un élément de procédure" whenTagName="ch-3-1" choice="single" perRegion="true">

    <Header value="Cible ?" underline="true"/>

    <Choice value="Affaire jugée"/>

    <Choice value="Autre affaire"/>

 </Choices>

      <Choices name="ch-3-1-3" toName="text" visibleWhen="choice-selected" whenChoiceValue="Problème juridique" whenTagName="ch-3-1" choice="single" perRegion="true">

    <Header value="Cible ?" underline="true"/>

    <Choice value="Affaire jugée"/>

    <Choice value="Autre affaire"/>

 </Choices>

      <Choices name="ch-3-1-1-1" toName="text" visibleWhen="choice-selected" whenTagName="ch-3-1-1" whenChoiceValue="Affaire jugée" perRegion="true" choice="single">

            <Header value="Auteur ?" underline="true"/>

            <Choice value="Du requérant"/>

            <Choice value="Du défendeur"/>
```

```
                <Choice value="D'un juge auteur d'une opinion séparée"/>
 </Choices>
  </View>
<!— la catégorie Résolution—>
<View visibleWhen="region-selected" whenLabelValue="Résolution">
   <Choices name="ch-4" showInline="true" toName="text" perRegion="true" choice="single" whenLabelValue="Résolution" required="true" requiredMessage="choix fonction rhétorique obligatoire">
                <Header value="Fonction rhétorique ?" underline="true"/>
                <Choice value="Donner des consignes aux juridictions compétentes"/>
                <Choice value="Rendre la conclusion de la Cour"/>
                <Choice value="Evaluer l'impact de la décision finale"/>
 </Choices>
</View>
<!— la catégorie fonction d'annonce—>
<View visibleWhen="region-selected" whenLabelValue="Fonction d'annonce">
   <Choices name="ch-5" showInline="true" toName="text" perRegion="true" choice="single" whenLabelValue="Fonction d'annonce" required="true" requiredMessage="choix fonction rhétorique obligatoire">
                <Header value="Fonction rhétorique ?" underline="true"/>
                <Choice value="Fonction d'annonce" selected="true"/>
 </Choices>
</View>
</View>
<!— chargement du texte —>
<View style="height: 700px; overflow: auto;flex: 70%;">
   <HyperText name="text" granularity="word" value="$text" valueType="text"/>
 </View>
  </View>    "
```

E. Iterative protocol to verify the annotation guide

1. The experts separately annotate a text from the reference sample using the current guide at the time.
2. The experts compare their annotations.
3. The experts agree on how to annotate divergent annotations, thus establishing a ground truth for the text.
4. The experts modify the guide to align it with the ground truth.
5. The experts send the guide to the testers, indicating the modifications, and have them annotate the same

text from the reference sample.

6. The experts modify the guide if the testers still make major errors or feel uncertain about how to interpret the guide.
7. Repeat steps 1-6 on a new text from the reference sample
8. Continue until the guide is no longer modified.
9. Once the guide is stabilized, the experts modify previously annotated texts from reference sample to make them coherent with latest version of guide and annotate the remaining texts from the reference sample.

F. Description of the training of external annotators

The theoretical training aimed to help external annotators understand move analysis and its significance for Lexhnology. Emphasis was also put on understanding the coding scheme which described the discourse units to be annotated, the steps, and the annotation guide. Practical training was then progressively conducted in a manner following the top-down approach (starting with the selection of a category and moving to the selection of attributes). The two annotators were also provided with an annotation video of a majority opinion, as well as with a decision tree to assist in choosing categories, rhetorical functions, and attributes. The annotators' performance was assessed through quizzes about the content of the guide and the annotation of five majority opinions from the reference sample.