Section-Type Constraints on the Choice of Linguistic Mechanisms in Research Articles: A Corpus-Based Approach

Inauguraldissertation

zur Erlangung des Grades eines Doktors der Philosophie

im Fachbereich Neuere Philologien

der Johann Wolfgang Goethe-Universität

zu Frankfurt am Main

vorgelegt von

Iverina Ivanova

aus: Veliko Tarnovo, Bulgarien

2021

Date of Disputation: September 24th, 2021

Acknowledgments

I would like to express my sincere gratitude to several people who contributed greatly in various ways to my taking up this challenge and embarking on this fascinating journey. These are the people whose unconditional support, sound judgments, and constant motivation encouraged me to dive fearlessly into the ocean of discourse analysis and inspired me to learn every day something new that helped me swim better. These are the people who taught me to set clear goals, guided me through my deep dives, and made sure that I come back unscathed. Sincere thanks to Prof. Dr. Gert Webelhuth, Prof. Dr. Frank Richter, Dr. Niko Schenk, Dr. Janina Radó, Prof. Dr. Manfred Sailer, Prof. Dr. Dianne Jonas, Michael Walzer, my mom and dad, and my brother.

Thank you all!

Abstract

This thesis investigates the structure of research articles in the field of Computational Linguistics with the goal of establishing that a set of distinctive linguistic features is associated with each section type. The empirical results of the study are derived from the quantitative and qualitative evaluation of research articles from the ACL Anthology Corpus. More than 20,000 articles were analyzed for the purpose of retrieving the target section types and extracting the predefined set of linguistic features from them. Approximately 1,100 articles were found to contain all of the following five section types: abstract, introduction, related work, discussion, and conclusion. These were chosen for the purpose of comparing the frequency of occurrence of the linguistic features across the section types. Making use of frameworks for Natural Language Processing, the Stanford CoreNLP Module, and the Python library SpaCy, as well as scripts created by the author, the frequency scores of the features were retrieved and analyzed with state-of-the-art statistical techniques.

The results show that each section type possesses an individual profile of linguistic features which are associated with it more or less strongly. These section-feature associations are shown to be derivable from the hypothesized purpose of each section type.

Overall, the findings reported in this thesis provide insights into the writing strategies that authors employ so that the overall goal of the research paper is achieved.

The results of the thesis can find implementation in new state-of-the-art applications that assist academic writing and its evaluation in a way that provides the user with a more sophisticated, empirically based feedback on the relationship between linguistic mechanisms and text type. In addition, the potential of the identification of text-type specific linguistic characteristics (a text-feature mapping) can contribute to the development of more robust language-based models for disinformation detection.

Table of Contents

List of Tables

List of Figures

Chapter 1 Introduction and Motivation

The current research aims to identify distinct linguistic features for the different section types that appear in the same research article, as well as to provide a quantitative and qualitative evaluation of the distinctions with regard to the communicative goals of the section types.

This research is also a natural continuation of a pilot project for the present work which sought to verify that the nature of the question and the communicative goal of the text type impose constraints on the choice of linguistic expression. The project and its results are described in Ivanova (2020). Building on this research, I, therefore, set the scope of this work again in the context of von Stutterheim & Klein (1989), who claim that each text is produced as an answer to a question and that the question influences the choice of information that is retained throughout the text and its realizations, the choice of cohesive devices that mark relations between entities both within and beyond the sentence boundaries, and the choice of rhetorical devices and their functions. In addition, Grosz & Sidner (1986) propose a framework for the analysis of text structure and the processing of utterances in discourse. They argue that a discourse consists of three different but constantly interacting structures: *linguistic* (consisting of segments in which utterances naturally aggregate), *intentional* (the author's intentions that the individual segments express and the relations among them), and *attentional* (the piece of information in focus at any given point in the discourse). The two frameworks will be analyzed in greater detail in Chapter 2.

The present study focuses on the interaction between the linguistic and the intentional structures of five section types -- *abstracts, introductions, related works, discussions,* and *conclusions.* To verify that the intentions of the individual section types constrain the choice of linguistic mechanisms, I investigate across these sections the frequency of occurrence of linguistic features associated with academic writing such as *nominalization, hedging, self-mentions*, *passive voice*, as well as features contributing to text coherence such as *coreference, lexical repetitions,* and *explicit connectives*.

The five constituent sections forming the article structure express distinct discourse intentions, i.e., the authors aim to achieve different communicative goals with them. *Abstracts*, for example, are aimed at persuading the target reader to read the full scientific paper by informing them about the purpose, the methods, the results, and the possible contribution of the conducted research in a succinct manner. *Introductions* describe the scope of the study -- its aim and how this aim can be achieved. In the context of computational linguistics, in introductions, authors tend to either test/compare the performance of different models to see which one performs better under particular circumstances, present a new model and describe its functionality, or present a different approach for solving a research problem. In introductions, one can also find a brief overview of the organization of the article's content. *Related works* relate the topic under discussion to previous works and studies. In this section, authors typically present the ways in which the current study differs from previous ones and emphasize the contribution of the present research in relation to what has already been investigated, i.e., in what ways the research adds to/expands some previous knowledge or differs from previous studies in terms of the adopted approach. *Discussions* present the results of the conducted research and the authors' subjective interpretation of these results. *Conclusions* provide a summary of the research procedure, the results, the claims that authors make based on the observations, their beliefs in the possible contribution of their study to the field of research, and their plans on how to extend or improve their results/the functionality of a model.

However different the intentions of these text types are, they complement each other since they can be considered intermediate answers to subquestions that add to the answer to the main question, namely *What is the research article about?*

These findings can, on the one hand, give us insights into the section-specific linguistic characteristics and, on the other, contribute to a better understanding of how the different sections interact with each other so that the overall purpose of the scientific paper is achieved. In addition, the results can help us deepen our knowledge of the rhetorical functions of the extracted features, as well as improve our understanding of the argumentation strategies employed in academic writing.

A Roadmap for this Dissertation:

In Chapter 2, I address the notions of discourse and discourse analysis, then I give a brief overview of the different approaches to discourse analysis, and finally, I analyze in greater detail those approaches that primarily inspired the current study. In Chapter 3, I provide details about the corpus used for the experiment and the linguistic features extracted from the target section types constituting the corpus. For each linguistic feature, I provide a brief description, motivation for its selection, and details about how it was measured. In Chapter 4, I make a quantitative assessment of the frequency of occurrence of each linguistic mechanism across the section types and offer an interpretation of the possible constraints that each section type might place on the frequency of occurrence and the rhetorical function of a linguistic mechanism. In Chapter 5, I elicit the linguistic mechanisms with which each section type can be associated. Then, I comment on the possible applications of such feature-section mapping. In Chapter 6, I summarize the results of the research and discuss possible future extensions of the study.

Chapter 2 Research Background

The object of analysis in the current study is the research article, which is a type of written discourse set in the academic context. In order to analyze the structure of this type of discourse, the distinctive linguistic mechanisms of its constituent sections, and the factors that might account for their frequency of occurrence, it is worth clarifying, first, two notions: *discourse* and *discourse analysis*. Then, I provide a brief overview of the approaches to linguistic analysis of discourse, which were designed in the context of various domains of research such as sociology, anthropology, philosophy, and linguistics, and which find application in problem-solving methods in disciplines such as psychotherapy, education, academic writing, artificial intelligence, etc. (Schiffrin, 1994). Afterward, I address the two frameworks on which the current study was based, namely, the Question Under Discussion (QUD), whose proponents are von Stutterheim & Klein (1989), and the framework proposed by Grosz & Sidner (1986), which stresses the role of discourse intention in the processing of utterances that aggregate in the discourse structure. Finally, I review some major studies exploring the characteristics of academic writing.

There are three major views of **discourse** and **the goals of discourse analysis**: *formalist (structuralist)*, *functionalist,* and a view seen as an intersection of the first two. The formalist perspective on language defines it as a unit of language above the sentence or above the clause. In fact, Harris (1952), the first discourse researcher to introduce the notion of discourse analysis, states that discourse is “the next level of hierarchy of morphemes, clauses, and sentences” and the goal of discourse analysis is to help language users understand what makes a discourse different from a random sequence of sentences. What is more, a central point of the structuralist views is that discourse is made up of smaller linguistic units (constituents) that are interdependent and that are arranged in a rule-governed manner (Schiffrin 1994, p. 24). These constituents can be morphemes that combine into sentences (Harris, 1988), clauses (Linde & Labov, 1975), or propositions (Mann & Thompson, 1988). Such a structure-based analysis of discourse suggests that the characteristics of its building blocks can provide insights into the discourse characteristics. However, Schiffrin (1994, pp. 28-29) points out two major pitfalls of such reasoning. First, she notes that this kind of analysis can lead to a circular definition of discourse in the sense that the properties of its constituents can contribute to or result from the properties of discourse and its purpose. Second, she argues that discourse structures are not always hierarchically organized, i.e., discourses are not always made up of morphemes that combine into words, words into phrases, phrases into clauses, and clauses into sentences. An example of which is spoken language in which exchanges are not always composed of full sentences.

The functionalist perspective on language, by contrast, posits that the form-based characteristics of discourse are not sufficient for its analysis and that there are extralinguistic factors such as context, speaker's intentions, and sociocultural aspects that also come into play when the discourse message should be interpreted. Hence, functionalists define discourse as language in use, and discourse analysis as "the study of any aspect of language use" (Fasold 1990, p. 65). In an earlier study, Brown & Yule (1983) also noted that the description of language forms in discourse analysis is dependent on the purposes and functions they were intended to serve in a communicative situation. Thus, discourse can be defined as a system through which particular functions are realized (Schiffrin, 1994).

Examples of language in use include dialogues, podcasts, social media posts, Zoom conference meetings, newspaper articles, argumentative essays, etc. As a communicative act, discourse involves participants: an initiator, i.e., a speaker or a writer/author (addresser), and also other participants -- hearers/readers (recipients/addressees) at whom the initiated form of discourse is aimed.

In addition, a discourse is placed in a particular context: sociocultural, political, medical, environmental, financial, and therefore, the interpretation of the meaning of the discourse is shaped by the context in which the discourse was constructed. Grimshaw (1981) and Foucault (1982) advocate the idea that discourse is the intersection of language and society and that both systems are interdependent and constitute each other.

What is more, a discourse has a particular intention (a discourse purpose), which means that the addresser comes up with a sequence of thoughts and directs them to their target recipient(s) to achieve a particular communicative goal such as to inform them about a topic/news/current research results, to convey their stance/evaluation of a situation, to express sentiment in response to someone's behavior/actions (e.g., happiness, anger, disappointment, fear, etc.), to impress, to entertain, to warn, to motivate, etc. Depending on the communicative goal of the discourse, the common ground (Stalnaker, 2002), and the context in which the discourse is produced, the addresser makes a series of linguistic choices regarding vocabulary, word order, intonation (in speech), information structure to get their message across. According to Grosz & Sidner (1986), a discourse purpose is achieved when the recipients have recognized the addresser's intentions unless it was the addresser's original intention that his/her purpose is not recognized.

Schiffrin (1994) proposed a third definition of discourse, which strikes a balance between the structuralist and the functionalist views. She defines discourse as utterances; a contextualized sequence of units of language in use. This definition captures, on the one hand, the formalist idea that discourse is a unit larger than single sentences, and on the other hand, it captures the functionalist idea that its constituents, the utterances, are contextually bound.

Approaches to discourse analysis such as speech act theory, interactional sociolinguistics, the ethnography of communication, and variation analysis, to name just a few, focus either on the structural view of discourse by identifying units, discovering regularities of underlying combinations, making decisions on whether certain unit arrangements are well-formed or on the functionalist view by emphasizing the importance of the contextual conditions such as the setting, the psychological state of the speaker (thoughts, feelings, intentions), as well as the sociocultural background and the interpersonal context for the successful interpretation/understanding of the function of the utterances.

According to speech act theory, which is based on the insights of the philosopher John Austin (1962), utterances do not simply describe or report states of affairs, but they can also be used to perform acts. Austin suggests that three speech acts underlie the production of an utterance: the *locutionary act* (i.e., the saying of a sequence of words, which have meanings), the *illocutionary act* (i.e., the addresser's communicative intentions; what is done by uttering the sequence of words), and the *perlocutionary act* (i.e., the consequential effect on the thoughts, feelings, or actions of the addressee). Austin asserts that all utterances are used to perform speech acts and have an illocutionary force, regardless of whether this act is linguistically explicit, i.e., it contains a performative verb like *promise, pronounce, name,* etc., or not. Thus, the act of stating, for instance, is viewed as an illocutionary act, as much as the act of promising, warning, or pronouncing. Austin also argues that the interpretation of illocutions is contextually bound. When participants take the speech situation into account, this allows them to separate one function of the utterance from another as in the example *Can you pass me the bottle of water?* If this utterance occurs in a medical setting, a doctor may ask a patient, who is suffering from a neuromuscular disorder, whether he/she is able to do the action. The utterance can be interpreted as a real question to which the doctor would expect the patient's answer-- yes or no.

This answer, ideally, should be accompanied by the performance of the action as a validation that the patient can perform the physical act. Another interpretation, which is the more common one, is to interpret it as a request that indirectly asks the addressee to perform the action. In this case, the addresser would not expect an explicit response with yes or no, but rather, a particular reaction on the part of the addressee (i.e., the passing of the bottle).

What is more, Austin argues that the truth and falsity of utterances are also dependent on context. For example, if the addresser makes a promise to be at a particular place, *I promise to be there*, the recipient would expect that the addresser would keep his/her word and be at the stipulated place. The procedure of promising would be true or felicitous if the addresser sticks to his/her promise. However, if the addresser utters the promise without intending to be at the stipulated place, this would abuse the procedure of promising. Contextual conditions such as the addresser's thoughts, feelings, and intentions are said to be part of the circumstances that make an utterance truthful or insincere (Schiffrin 1994, pp.52-53).

In 1969, the philosopher John Searle proposed a framework that incorporated Austin's speech acts into speech act theory, which offers an approach to discourse analysis. According to this approach, a discourse consists of units (utterances) that have communicative functions that can be identified and labeled. These speech acts can be governed by various contextual and textual conditions, and their significance for the discourse analysis is that they can respond to and at the same time initiate further acts (Schiffrin 1994, p. 91). The recognition of the communicative function of an utterance enables the specification of the type of response, i.e., it creates options for the following utterance. However, there are cases in which an utterance can have multiple illocutions/multiple functions, as was illustrated above, and it is the contextual circumstances that constrain the options for subsequent utterances.

The **interactional sociolinguistic approach** views discourse as a sociocultural construct; a unit created out of the interaction between language, society, and culture, and their relationship with the self and the other. The approach suggests that language in discourse is molded in a way that mirrors the social and interpersonal contexts, which enables recipients to draw meaningful interpretations from communicative situations based on the social prerequisites. This analysis of discourse was inspired by the insights of the anthropologist Gumperz (1982) and the sociologist Goffman (1963) who noted that the meaning of language can be inferred from social and interpersonal contexts[1]. Goffman, for instance, suggested that all utterances in a discourse

[1] The following exposition is based on Schiffrin (1994).

are contextualized, i.e., are situated in occasions, situations, and encounters that determine not only its structure but also facilitate the interpretation. Central to this approach is, therefore, the idea that the contextualization of each utterance in a discourse motivates its use (Schiffrin 1994, pp. 133-134).

Another approach to discourse analysis, grounded in anthropology and linguistics, is **the ethnography of communication**. This approach focuses on the relationship between cultural beliefs/norms and language use. In other words, it studies how the social expectations in certain speech situations, or the cultural norms/beliefs influence the way utterances are constructed and used (Schiffrin 1994, p. 185). It also provides a framework for the analysis of the interaction between speech acts and speech events. Schiffrin applied the approach to investigate this type of relation. To be more specific, she investigated how questions (as examples of speech acts) fit into the structure of interviews (as examples of speech events). To do so, she compared the types of questions that occur in two types of interviews and found that the form and the function of the questions are dependent on the communicative structure of the interviews. Schiffrin pointed out that information-seeking questions, for instance, might not require prior information or may not be semantically dependent on previous utterances. However, questions asking for confirmation or clarification are dependent on prior context. The approach, therefore, posits that the continuity between the utterances is constrained by the nature of the speech event in which the speech acts occur (Schiffrin 1994, pp. 184-185). What one would expect in a particular type of discourse would influence the way the utterances are formed and the way they interact with each other. Similarly to speech act theory, the ethnography of communication focuses on structures made up of sequential utterances in which each utterance is labeled with the speech act it performs derived from its linguistic characteristics (Schiffrin 1994, p. 356).

Finally, **variation analysis** is concerned with how the surface features of utterances and the acts these utterances are used to perform build structure across clauses. By analyzing different text types such as narratives, lists, etc., Schiffrin found that surface features of utterances make a text coherent when they realize certain functions associated with the text type at a particular location within the text. She also points out that these surface features can be treated as variables, i.e., their realizations tend to be dependent on their location in the text (Schiffrin 1994, p. 355). The approach that I adopt for the analysis of the section types bears some resemblance to variation analysis in the sense that I seek to explain how the assumed purpose of the text (the section type) influences its surface linguistic features and their functions. An example of this would be whether the frequency of hedges and their rhetorical functions can be influenced by the communicative goal of the section type.

In the present research, I investigate the effect of the addresser's intentions in language use in academic discourse rather than the effect of the social and cultural aspects (e.g., author's gender, author's mother tongue, year of publication, etc.), which can also shape the discourse characteristics.

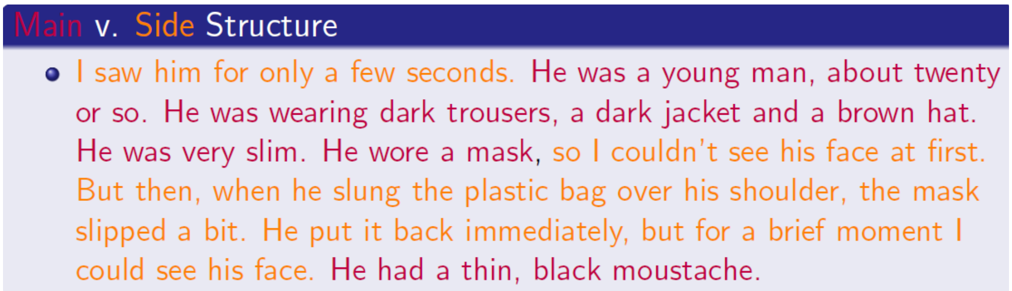
The research is based rather on two frameworks of discourse analysis: the Question Under Discussion (QUD), and Grosz & Sidner's approach to discourse analysis based on the theory of attention, intention, and discourse structure. These frameworks seek to explain how the individual section intentions and the question that each section in a research article answers influence the choice of linguistic mechanisms.

The QUD approach (Benz & Jasinskaja, 2017) views discourse as an answer to a question (Quaestio/the Big Question), and all the utterances that aggregate in the discourse structure either directly answer the Queastio or answer subquestions that contribute to answering the Quaestio. This question-answer approach seeks to characterize how a sentence fits in the context of the discourse.

von Stutterheim & Klein (1989) propose that the nature of the question can constrain the choice of information that appears in the discourse. In other words, the choice of semantic domains (referents) such as *time, place, persons/objects, events,* and *modality* is constrained by the text type and the question the text is produced to answer. This means that the question determines how the referents are packaged in the utterance structure, i.e., whether they function as the topic/given information or the focus/new information (Chafe, 1976), which of them are retained as the discourse unfolds, and what are their categorical realizations (von Stutterheim & Klein, 1989).

von Stutterheim & Klein (1989, pp. 46-47) refer to the ways in which information from the five domains shifts from one utterance to another as referential movements. These referential movements provide insights into the topic-focus structure of each utterance and contribute to the topic persistence and the connectivity of the utterances both on textual and idea levels.

To illustrate how the text intentions influence the choice of domains of reference, they investigate how the QUD determines if a sequence of utterances can be interpreted as a description or a narrative. Thus, for example, the text that will be produced as an answer to the question *How did the robber look?* will be a description of a robber. Therefore, the types of information that the recipient would expect to hear/read will be references to a person (the robber) and some properties (descriptions of his appearance) like in the example below, which is an excerpt from a witness account. According to von Stutterheim and Klein, the utterances, which directly answer the QUD, i.e., which provide the actual description, form the main structure of the discourse, and the rest of the utterances form the side structure, i.e., they provide background information, which is also relevant but not part of the description itself. For readability purposes, I have highlighted the main structure in red and the side structure - in orange.

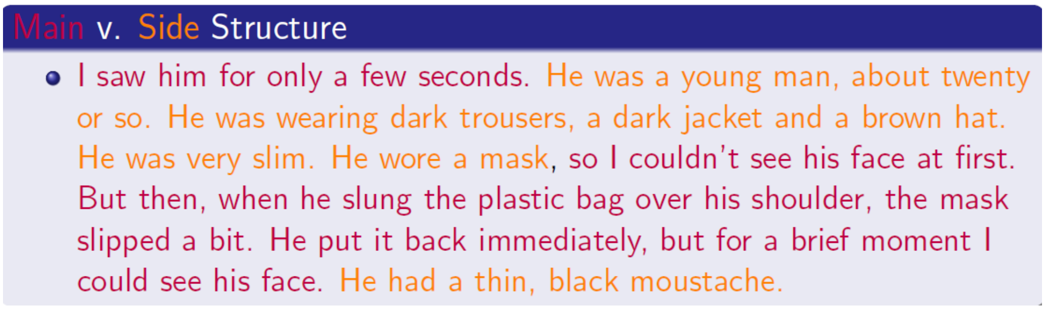


Main v. Side Structure

- I saw him for only a few seconds. He was a young man, about twenty or so. He was wearing dark trousers, a dark jacket and a brown hat. He was very slim. He wore a mask, so I couldn't see his face at first. But then, when he slung the plastic bag over his shoulder, the mask slipped a bit. He put it back immediately, but for a brief moment I could see his face. He had a thin, black moustache.

Figure 1 The utterances forming the main and the side structure in the excerpt when it answers the question "How did the robber look?" (von Stutterheim & Klein, 1989)

However, if the witness was asked the question *What happened?*, the utterances that provide a narrative of the event (i.e., what the narrator saw) would form the main structure of the discourse, and the utterances with the descriptions would form the side structure:



Main v. Side Structure

- I saw him for only a few seconds. He was a young man, about twenty or so. He was wearing dark trousers, a dark jacket and a brown hat. He was very slim. He wore a mask, so I couldn't see his face at first. But then, when he slung the plastic bag over his shoulder, the mask slipped a bit. He put it back immediately, but for a brief moment I could see his face. He had a thin, black moustache.

Figure 2 The utterances forming the main and the side structure in the excerpt when it answers the question "What happened?" (von Stutterheim & Klein, 1989)

von Stutterheim and Klein go on further to explain that the domains of reference are constrained by the two text types. When it comes to the **temporal domain**, they suggested that in narratives one would expect a specific time interval since the series of actions are normally presented in chronological order. In descriptions, however, time reference remains constant throughout the text. The **spatial domain**, similarly to the temporal domain in narratives, provides the sequential organization of the information, so one would expect frequent references to locations. In descriptions, references to locations may play a role if, for example, the addresser describes a picture and explains what objects they can see and where they are located in the picture. When it comes to the **persons/objects domain**, in narratives there is usually a reference to one specific person, the narrator, or a group of persons, known as the protagonists of the story. The simplest scenario is when there is one protagonist who takes up the subject position in a sentence. If the narrator introduces a new protagonist, he/she will be the focus and will occupy the object position in the sentence in which they were first mentioned. In the subsequent utterances, the newly introduced referent will become the topic, will occupy the subject position, and may be realized by a personal pronoun or a definite NP. In descriptions, the referent can be specific or non-specific depending on the nature of the question, and if the question places a person/object in focus, then there will be a different information packaging.

Predicates express events or properties that belong to the focus in the utterance structure. In narratives, they are realized by VPs describing actions, processes, or states that are temporally bound. In descriptions, predicates are again realized by VPs but semantically denote visual properties and descriptions ascribed to the person/object under discussion.

Finally, **modality** involves two major concepts. The first one describes the relationship between the validity of a proposition with regard to possible worlds, which gives rise to the following subtypes: necessity (when the proposition is valid in all possible worlds), possibility (true in at least one possible world), real (true in what is considered the "real" world), fictitious (true but not in the "real" world). The second concept is related to the addresser's opinion/ judgment/evaluation of the validity of a situation, or the degree of certainty with which they make a proposition about a situation. Modality is usually realized by modal auxiliary verbs (*might, may, could, should, etc.*), modal adverbs (*perhaps, possibly, certainly, fortunately, etc.*), or modal lexical verbs (*believe, think, desire, etc.*), and depending on the nature of the question, they may be employed to serve different functions, e.g., to express necessity, obligation, possibility, certainty, etc.

In narratives, it is suggested that there is a constraint on one type of modality, the modality denoting the validity of the proposition in both the "real" world or its validity in the "fictitious world" in which the story world is treated as if it were real. Modality belongs to the topic and the main structure. In descriptions, there is no constraint on the validity of the proposition in the real world or the world treated as real (von Stutterheim & Klein 1989, pp. 49-54).

In this dissertation, I investigate which of these domains of reference are activated under the effect of the section purpose by analyzing the semantics of the lexical repetitions.

What is more, van Kuppevelt (1995), as a supporter of QUD, also points out that discourses have a hierarchical structure of explicit or implicit questions, and the role of the questions is to determine what the discourse is about and what the topic of each utterance is.
Other proponents of the QUD approach such as Benz & Jasinskaja (2017) recognize and report on its successful application in the analysis of various linguistic phenomena such as presuppositions and implicatures, and how implicatures can be inferred from QUDs.

Roberts (1996) also adopts the QUD approach to explain the underlying principles of discourse creation. To do so, she pursues an analogy with games. Similarly to a game, a discourse has a goal (a set of goals) and rules to which discourse participants have to adhere in order to achieve the discourse goal. The rules can be conventional (syntactic, compositional semantic) and conversational (the Gricean maxims). There are also moves that participants make towards achieving the aim of the discourse, as well as strategies they employ to make these moves. The strategies, however, are constrained by the discourse goal, the rules, and the participants' moves. The moves are of two types: set-up moves (questions) and payoff moves (assertions/answers to questions). The question determines a set of propositions (the correct answers) or the proffered alternatives (the asserted alternatives). Once a question is accepted, it becomes the immediate question under discussion (the QUD) and the discourse participants commit to answering it. This common intention becomes part of the common ground of the participants. To answer the QUD, the participants abide by rules: *the maxim of relevance*, which contributes to text coherence, which, in turn, improves information processing, and *the maxim of quantity*, which makes participants provide a full, not partial, answer to the question. Since a discourse is viewed as an answer to a question (the Big Question), its goal would be to provide an answer to it. Participants, therefore, develop a plan (strategy) to achieve this by developing subgoals, i.e., answering subquestions that are more specific than the big question and more manageable. Such a strategy of attacking the big question would have a hierarchical structure in which questions are logically related by contextual entailments (Roberts 1996, pp. 92-96).

A theory, which is particularly influenced by Roberts' hierarchical discourse model, is Büring's (2003) pragmatic theory of contrastive topics, the conditions under which they occur, and their relations to foci. In order to analyze the contrastive topic-focus relation, he adopts the hierarchical discourse model and represents it in the form of discourse trees (d-trees). The matrix node (a discourse) is represented as a hierarchical structure of questions. These questions may consist of daughter subquestions whose answers may fully or partially answer the question under discussion (i.e., the d-topic). He points out that informativity and relevance are two principal constraints on constructing d-trees. Informativity involves the common ground, i.e., the knowledge that discourse participants share so that any redundant information is avoided. Relevance involves adherence to a question till it is resolved. Each node stands for a move and in order for a move to be relevant, it must answer or at least address the question under discussion. Büring emphasizes that such a d-tree representation of discourse is the basic representation of the hierarchical structure necessary for the interpretation of occurrences of contrastive topics and foci.

In light of the principles presented above, the current work focuses upon the research article as a type of discourse, which is also produced as an answer to a question, for example, *What is this article about?* To provide an answer to this question and to present the information to the research community effectively, authors adopt an organizational strategy that is intended to provide a systematic presentation of the research information and facilitate its processing. This strategy is the segmentation of the article into subsections. These subsections can be viewed as answers to subquestions such as *What is the focus of this article and how is its content organized?, What related work has been done?, What is the scope of the research and how does it differ from previous studies?, How was the study conducted?, What are its results?, How can the results be interpreted?* The constituent sections in a research article are organized in such a way that the content of one section naturally prepares for the content of the subsequent section. In simplest terms, one would expect to learn about the methodology of the research before being exposed to its results. Each section as an answer to a subquestion either provides a direct or a partial answer to the superquestion. I take this view of discourse as given and focus on the linguistic strategies that authors employ inside each of these sections to answer first the subquestions and eventually the superquestion.

In addition, the QUD is viewed as one of the fundamental building blocks of coherent text structures and therefore, it finds implementation in QUD-based models that seek to find a mapping between coherence relations and QUD-based structures. For instance, Ginzburg (2012) implements the approach to model the structure of a natural dialogue as a set of query-reply pairs. The questions under discussion at any given point of the dialogue contribute to narrowing down the possible topics that discourse participants can select from (i.e., what was said last, what to say next, and when to say it). Each question may evoke extra questions whose job it is to further clarify or trigger a more specific answer to the initially posed question, or sometimes trigger another topic.

He also suggests that QUD is an ordered set of questions and this can be motivated by the fact that more than one question can be under discussion at a given point, but there will be one question that tends to predominate. In other words, at any given point in the conversation, there will be a question that will deserve precedence over other questions (Ginzburg 2012, p. 68). He goes further to explain that a conversation would be coherent even if an interlocutor responds to a question with another question if they believe that the topic this question addresses deserves to be prioritized over the other questions, with the idea that answering this question would ensure the success of the communicative situation.

In certain situations, however, he suggests that the ordering of the QUD can be negotiated and that the occurrence of a particular question can be for organization purposes, i.e., to determine the order in which the answers to certain questions should be resolved. He illustrates this with the following example:

(1) *A: Who did Bill invite?*
B: Which of his friends do you know?
A: Before I can answer this, you really need to answer my question?
B: But I cannot answer it before you answer mine. (Ginzburg 2012, p. 70)

Ginzburg concludes that conversation structure is very much participant-intrinsic and dependent on their mutual agreement to discuss a question (topic), and so is the ordering of the QUD.

Another approach that seeks to explain the structure of a coherent discourse is proposed by Grosz & Sidner (1986). According to this approach, a coherent discourse consists of three distinct but interconnected structures -- linguistic, intentional, and attentional. Grosz and Sidner suggest that a discourse can be composed of more than one discourse segment in which utterances aggregate. The linguistic structure is made up of segments/units of language, which in turn are made up of a sequence of utterances. The relations between the segments are marked by surface linguistic devices (cue phrases). The segment structure and the linguistic cues are said to be interdependent, i.e., the linguistic devices can provide insights into the structure of the individual segments, and the structure of the individual segments can constrain the function and the interpretation of the linguistic devices. What is more, the linguistic devices can serve as explicit indicators of a change in the segment's intention or a change in focus.

The intentional structure captures the intentions associated with the individual discourse segments. Grosz and Sidner explain that the recognition of the purposes of the component segments not only characterizes their relevance and contributes to the recognition of the purpose of the overall discourse, but also allows discourse participants to distinguish a coherent from an incoherent text. Some intentions are meant to be recognized, others might remain private. The intention/purpose of a segment is achieved if the recipient recognizes its underlying intention. A segment might have a range of intentions but what is critical for the analysis of discourse structure is for the discourse participants to identify the relevant relations that hold between the intentions of the segments (Grosz & Sidner 1986, p. 179).

The attentional structure records the objects/properties/relations which are relevant or in focus at any point as the discourse unfolds. The attentional structure can be related to the referential movements constrained by the question the discourse answers and the addresser's intentions. The role of the attentional structure is to keep track of the information already activated in the previous utterances, which is important for the interpretation of the subsequent utterances (Grosz & Sidner 1986, p. 177).

This theory of discourse provides solid foundations for investigating the structure and the meaning of various discourse types. It posits that a discourse is coherent when each utterance contributes directly or indirectly to achieving the discourse purpose. In addition, it suggests that the change in intentions would influence the choice of attentional structures, and this, in turn, would constrain the choice and the frequency of the surface linguistic devices.

In the current study, I investigate the interaction between the linguistic and the intentional structures of the sections constituting the structure of a research article and attempt to provide quantitative verification of how the hypothesized intentions of the individual sections constrain the types of referents that are activated and the types of linguistic devices that mark the relations between meanings inside each section and between the sections. I address the attentional structure only in relation to identifying the topics/the objects of attention that are retained throughout the article and the topics that are activated only in particular sections (i.e., section-specific topics/keywords).

Since the analyzed discourse segments are set in an academic context, it makes sense that I narrow down my analysis to the linguistic features that are considered typical of academic writing such as nominalizations, hedges, passive voice constructions, objective language, and self-mentions. In addition to them, I also examine if the purposes of the sections influence the frequency of cohesive markers such as coreferences, lexical repetitions, and connectives.

Previous studies focus on the identification of the linguistic features that set scientific and non-scientific language apart. Their results show that the extensive use of complex noun phrases, specialized vocabulary, and the use of passive voice, which is believed to contribute to the objective transmission of knowledge, can be considered an integral part of the scientific expression (Veel, 1997; Wellington & Osborne, 2001; Schleppegrell, 2004; Ahmad, 2012). Other scholars focus on academic discourses such as research articles and argue that scientific language can also convey the author's stance and beliefs. Hyland (2001), for example, compares the frequency of occurrence of self-mentions in research articles coming from the humanities and the natural sciences. The result of his study confirms that authors make self-references and they do so more frequently in the humanities than in the natural sciences. Yazhilarda et al. (2017) also investigate the frequency and the rhetorical functions of self-references in research articles. Contrary to the results of earlier studies on scientific language, which posit that it is devoid of personalized expression, Yazhilarda et al., for example, emphasize that the use of personal pronouns is central to scientific writing. In fact, they argue that authors employ them to achieve various rhetorical functions such as to present the aim of the conducted research, to explain the procedure, to discuss its results, and to provide their subjective interpretation of the presented results.

Other studies compare the scientific writing styles of native and non-native English language scholars by analyzing linguistic complexity in two dimensions: *syntactic complexity* measured by sentence length and sentence complexity, and *lexical complexity* measured by lexical diversity, lexical sophistication, and lexical density (Lu et al., 2018). The results from the study show that native speakers of English tend to produce longer sentences containing more clauses, make use of longer nouns and shorter verbs, and utilize more verbs and fewer nouns (Lu et al. 2018, p. 27).

What is more, Hyland (1998) investigates the distribution of hedging types across the section types in research articles and provides a detailed account of their distribution in each section, their realizations, and the rhetorical functions they can have.

Other studies concentrate on the grammatical and lexical characteristics of scientific texts to measure text coherence. For example, Wang & Zhang (2019) analyze the occurrence and the distribution of lexical cohesive devices across the section types in research articles coming from different academic disciplines. More specifically, they analyze how the intentions of the sections influence the frequency of occurrence of particular lexical devices such as lexical repetitions, synonyms, antonyms, meronyms, etc.

The research that I conducted bears a resemblance to the studies mentioned above in the sense that it analyzes the linguistic characteristics of academic writing and explores their occurrence and distribution in research articles. However, it also differs from them in three major ways: first, it focuses on research articles and their constituent sections in one particular scientific discipline -- Computational Linguistics (CL); second, it examines the effect of section type intentions on the choice of the combination of features associated with academic writing and features contributing to text coherence; third, based on the quantitative assessment of the results, it seeks to elicit distinctive section type characteristics.

Chapter 3 Experimental Setup

For the purpose of the current experiment, I analyzed more than 20,000 research articles in the field of computational linguistics (CL) available from the ACL Anthology Reference Corpus[2] in order to retrieve the target sections -- *abstracts, introductions, related works, discussions,* and *conclusions*.

3.1 Data Preprocessing

Those papers that did not contain all of the target sections were removed from the final version of the corpus. Moreover, not all scientific papers followed the same encoding models, so for some papers, only 2 or 3 sections were detected. It can be the case that these sections were not accurately detected or some sections were overlapping, e.g., in some articles, *introductions* and *related works* formed one section. The same holds true for *discussions* and *conclusions*. This may suggest that the conference papers in the corpus follow different schemes/guidelines for content organization, and therefore, all five sections could not be automatically retrieved from the whole corpus. Appendix 1 provides details about the NLP techniques/methods I used for the extraction of the five sections.

Since the present work aims to analyze the distribution of the linguistically-motivated features across all five section types appearing in an article, those articles in which one or more sections were absent were not considered for the quantitative analysis.

The frequency analysis of the linguistic features is, therefore, based on 1,119 research articles in which all five sections are present/detected (5,595 sections). Table 1 provides an overview of the average length of each section and its hypothesized intention.

[2] https://www.aclweb.org/anthology/ (Originally downloaded from: https://acl-arc.comp.nus.edu.sg/)

Table 1 An overview of the analyzed section types, their average token-based length, and their hypothesized intentions.

Section Type	Average # Tokens	Hypothesized Section Intention
Abstract	121.3	To inform about the purpose, the methods, the results, and the contribution of the research in a succinct manner. To convince the reader to read the whole research paper by emphasizing the significance of the research.
Introduction	653.0	To set the scope of the current research. To provide a roadmap for the paper's content.
Related Work	609.2	To relate the current study to previous ones. To emphasize the differences. To explain how the current study expands/ improves on previous results.
Discussion	711.6	To present the research results and evaluate them.
Conclusion	203.7	To make claims based on the results. To express beliefs/hopes about the research potential.

The section types were compared on the basis of the frequency of occurrence of predefined linguistic features. The source codes and their documentation are available on GitHub and also appear in Appendix 4. Any reference to GitHub is a reference to the following URL: https://github.com/iverinaivanova/Linguistic-Mechanisms-in-the-Section-Types-of-a-Research-Article.

The extracted section types for each research article, the source codes, the dataset, and the statistical analyses can also be downloaded from the Goethe University Data Repository (GUDe). DOI: 10.25716/gude.1jnt-32xh

3.2 Linguistic Features

The section types were compared on the basis of the frequency of occurrence of predefined linguistic features that are either associated with academic writing (Ahmad, 2012; Hyland, 1998) or are considered to improve text readability and overall text coherence (McNamara et al., 2009). Table 2 provides an overview of the features and their measurement. All features were automatically extracted with the help of frameworks for Natural Language Processing (NLP) which take raw texts as input, process them, and with the help of statistical models, tokenize them, assign parts of speech to the tokens, lemmatize, check if the token is a number, word, or a punctuation mark, identify the head word in each constituent and its dependent words, segment the texts into sentences, and assign syntactic dependency labels by describing the relations between the constituents in terms of their syntactic functions (e.g., subject, direct/indirect object, adjunct, etc.).

In the current section, I describe each analyzed linguistic feature, provide reasons for its selection, and explain briefly how it is measured.

Table 2 The linguistic features and their measurement.

	Feature	Measurement
1.	Nominalization: Noun Phrase (NP) Count	The total number of NP occurrences per text is normalized by the total number of tokens.
2.	Nominalization: NP Length	NP length is measured by counting the number of tokens per NP. Then all token-based numbers are summed up and the total is normalized by the total number of tokens per text.
3.	Nominalization: NP Complexity	The total number of NPs within which the head takes a dependent (Adj/Noun/PP/Past Part) is normalized by the total number of tokens.
4.	Self-mentions	The total number of self-referring words is

		normalized by the total number of NPs per text.
5.	Hedging: Passive Voice	The total number of passives is normalized by the total number of tokens.
6.	Hedging: Modal auxiliary verbs	The total number of modal auxiliaries is normalized by the total number of verbs.
7.	Hedging: Modal lexical verbs	Measurement based on Mutual Information (MI) scores (more on MI in Appendix 2)
8.	Evaluative and Objective Language	Measurement based on MI scores of content words (N, V, Adj, Adv)
9.	Cohesion: Coreference	The total number of coreference chains per text is normalized by the total number of sentences.
10.	Cohesion: Lexical Cohesion (Repetitions)	The total number of lexical chains marked by repeated lexical words is normalized by the total number of tokens.
11.	Cohesion: Explicit Connectives (Discourse Relation Types)	The total number of temporal/comparison/contingency/expansion connectives per text is normalized by the total number of tokens.

3.2.1 Nominalization

Nominalization is the process of forming nouns from other parts of speech, with or without the addition of inflections. In academic writing, this refers to the use of nouns instead of verbs. Such nominalizations are an integral part of academic discourse and are said to add to the technicality and abstraction of texts (Ahmad, 2012).

The current study investigates the distribution of nominalization instances across the five section types in order to verify, first, if sections have a different frequency of nominalization. Second, which are the sections that possess the most syntactically complex and longest NPs, and third, if there are differences, can they be motivated by the purpose of the section type.

The nominalization feature is analyzed in 3 dimensions: **NP Count**, **NP Length**, and **NP Complexity**. My hypothesis is that abstracts will have a low frequency of NP occurrences but will have the longest and most complex NPs. This can be a natural consequence if the small size of abstracts constrains writers to convey the central points of their research topic in a space-efficient manner, so they tend to condense this information into NP structures in which noun heads are heavily modified. In Ivanova (2020), for example, I found that in abstracts, there is a low frequency of NP occurrences and that the noun occurrences are more densely distributed than those in the article body, i.e., the distance between noun occurrences in abstracts is smaller than the distance between the noun occurrences in the body of the scientific paper. I argue that this can be motivated by the size and the purpose constraints that abstracts impose on the linguistic structure. If NPs are of lower frequency and nouns are densely distributed in abstracts, this can suggest that an NP structure in abstracts is complex and the noun head might take a set of noun modifiers, which augments the NP structure, makes it more informative, and at the same time more difficult to process. In the article body, by contrast, there are no such size constraints as the information can spread over several utterances instead of being condensed into a constituent below the level of the utterance.

The NP Count feature is measured by dividing the total number of NP occurrences by the total number of tokens. The **NP Length feature** is intended to verify two hypotheses, first, that section types differ in terms of the length of their NPs, and, second, that NPs in abstracts are longer token-wise than those in any other section type. The NP Length is measured by counting the number of tokens per NP. Then all token-based numbers are summed up and the total is normalized by the total number of tokens per text.

The **NP Complexity** feature investigates the frequency of NPs in which the noun head takes dependents such as adjectives (Adj), nouns (NOUN), preposition phrases (PP), or past participial clauses (VBN). The total number of NPs containing a dependent is normalized by the total number of tokens. This feature is intended to verify that abstracts contain the most syntactically complex NPs in comparison to the rest of the section types.

3.2.2 Self-Mentions

Academic texts are not simply a sequence of impersonal statements conveying scholarly knowledge to the research community, they are also a tool for the writers to define their identity and emphasize their contribution to the research field by explicitly referring to themselves. In fact, the role of such self-references (self-mentions) as a significant rhetorical marker has already been acknowledged by researchers interested in the characteristics of academic writing. Hyland (2001), for example, suggests that academic language is not devoid of the authors' voice. Authors tend to refer to themselves when they formulate their research aim, when they explain the research procedure, when they make claims based on their research results, and when they elaborate on an argument (Hyland 2001, p. 257).

He also proposes that the frequency of occurrence of such identity manifestations may vary from discipline to discipline and that authors may employ them to achieve different rhetorical goals. By comparing 240 research articles coming from hard and soft sciences, he found that the frequency of occurrence of first-person pronouns (singular or plural) is higher in soft sciences than in hard sciences and that in soft sciences writers refer to themselves mostly to express their personal stance or to make claims based on their findings, whereas in hard sciences they do so mostly to present the research methodology/procedures.

The current research is an extension of Hyland's study in the sense that it attempts to investigate not only the frequency of occurrence of self-mentions in CL research articles and to identify their rhetorical functions, but also to analyze their distribution across the individual section types that constitute the article. The results can shed some light on the section-specific frequency of self-references, as well as the possible rhetorical functions these can have with regard to the intentions of the section types. The frequency of self-mentions has been measured by dividing the total number of self-referring words per article section by the total number of NPs.

My hypothesis is that authors will refer to themselves most frequently in abstracts, discussions, and conclusions. In abstracts, they tend to present briefly their research aim, the approach they adopted, and their research findings. In discussions, they usually elaborate on the series of methods they used in order to achieve the reported results, and in conclusions, they tend to make their claims based on the presented results and express their personal stance on the significance of their research contribution, and the usability of their findings in future studies.

3.2.3 Hedging

Another central feature associated with academic writing is hedging. The notion of hedging was first introduced by Lakoff (1973) to refer to all linguistic devices that make an author's expression fuzzier or less fuzzy in speech and writing. There is a whole plethora of such devices signaling fuzzy language. For example, the use of modal auxiliaries like *might, could, may*, etc., or modal adverbs like *possibly, probably, almost,* etc., as well as some impersonal constructions such as *It appears to be...*, *There seems to be...,* etc. Since academic writing is characterized as a formal, objective, highly technical presentation of some piece of knowledge, one would not expect the use of words that confuse or provoke doubts in readers. However, previous studies, which focus on the pragmatic significance of hedging as a key element to effective argumentation in academic writing, have shown that hedging can have a variety of functions that enable authors to modulate their tone of expression, i.e., allow them to present their claims "with precision, caution, and modesty" (Hyland, 1998). Hyland reports that the employment of such devices strengthens, on the one hand, the credibility of the writer's statements and diminishes, on the other, the degree of criticism/rejection on the part of the recipient (Hübler, 1986). Hyland also proposes that hedges can signal either authors' lack of complete commitment to a proposition or their refusal to convey categorical commitment (Hyland 1998, p. 2).

Depending on the functions of hedges and their linguistic realizations, Hyland distinguishes two major types of hedges: content-motivated and reader-motivated (Hyland, 1998). The content-motivated hedges are concerned with the manner in which writers present and interpret the status of the proposed piece of knowledge. This type of hedge can be further divided into two contrasting subtypes -- the writer-motivated and the accuracy-motivated hedges. Writer-motivated hedges are believed to hide the writer's presence in the text, whereas accuracy-motivated hedges are said to imply that the claims writers make are based on inferences rather than on facts. Writer-motivated hedges, which according to Hyland are often realized by impersonal constructions and passive voice, reflect the writer's lack of commitment or the writer’s attempt to avoid making any categorical commitments to the propositions made by other scholars. By using such impersonal constructions, writers make statements with greater caution since they are not certain about the correctness of the claims and thereby, attribute the belief in the credibility of the claims to the scholars who have made them. Thus, by limiting their personal liability to the assertions, writers tend to reduce the risk of being judged for conveying other researchers’ claims as truths/categorical commitments. Impersonal expression can, therefore, be viewed as a central strategy of effective scientific argumentation. By contrast, the accuracy-motivated hedges can enable writers to convey as accurately as possible the probable interpretations of the claims based on their logical reasoning, derived from the content they have been presented with, or to convey the degree of certainty with which they interpret the statements made by other scholars. The use of such hedges can be viewed as an essential indicator of the writer’s critical thinking and of their competence to regulate their tone of commitment depending on the degree of expertise they have in the subject matter. According to Hyland, these hedges are mostly realized by epistemic modals such as modal auxiliaries (e.g. *might, may, should, could, etc.*), modal adjectives (e.g. *possible, probable, etc.*), adverbs (e.g. *possibly, perhaps, maybe, etc.*), and lexical verbs (e.g. *suggest, seem, appear, believe, hope, etc.*).

Last but not least, the reader-motivated hedges can reflect the writer’s attempt to improve writer-reader interaction by allowing authors to present their own claims not as facts but rather as personal conclusions (Hyland, 1998). These hedges are believed to signal that the writer’s claims are only one possible interpretation of the research results or one possible perspective from which the research problem has been addressed. By using such hedges, writers implicitly invite their target readers to participate in the dialogue, comment, criticize, and/or suggest alternative interpretations. What is more, Hyland suggests that these hedges have the effect of increasing the credibility of the proposed statements and decreasing the degree of rejection. In his study, he found that they are realized either by self-mentions or by epistemic modality. The use of such linguistic devices is said to emphasize the degree of caution and modesty with which authors formulate their claims. Authors acknowledge the fact that the assertions they make are based only on limited data, so the use of modals tones down the degree of confidence and certainty with which they present these claims. In addition, by referring to themselves, authors might emphasize the fact that the claims are based on their personal opinion or that they reflect an alternative evaluation based on their own findings.

In the present study, I seek to analyze the distribution of hedging occurrences across the section types and to determine the predominant type of hedges in each section on the basis of its realizations by adopting Hyland's categorization and the list of linguistic realizations he proposes. The results can provide some insights into the rhetorical functions of hedges in the section types of a research article in the field of Computational Linguistics, as well as deepen and improve our understanding of the argumentation strategies employed in academic writing. I measure hedging in two dimensions depending on the linguistic devices that realize it. These dimensions include hedging marked by passive voice and by epistemic modality, which is signaled by the occurrence of modal auxiliary verbs and modal lexical (non-factive) verbs. The frequency of passive voice has been measured by dividing the total number of passive voice constructions by the total number of tokens per text. Hedging marked by modal auxiliaries has been measured by dividing the total number of modal auxiliary verb occurrences by the total number of verbs. The frequency of non-factive verbs is based on Mutual Information scores. The computed values are available on GitHub and on GUDe under **Linguistic Mechanisms > MI scores > Modal Lexical Verbs**.

My initial expectation is that section types will differ in terms of **the frequency of the occurrence of hedges**, **the realizations of the hedges,** and **the rhetorical functions of these hedges**. Considering the purpose of the individual section types and the hedging markers, my hypothesis is that passive voice as an impersonal construction will be more frequent in introductions and related works. In introductions, authors often address the research problem and present an outline of all activities that have been performed in order for this problem to be solved or a particular model/approach to be evaluated without focusing on the performer of the activities. In related works, authors normally make references to previous studies by summarizing the methods their peer scholars have adopted and the observations they have made. Therefore, authors would detach themselves from the presented content in order to indicate that they were not involved directly in the described processes and that they cannot fully commit to the correctness of the findings presented in previous studies. In addition, authors may also mark the degree of certainty with which they interpret the results of previous studies by using epistemic modals.

Since abstracts are intended to promote and persuade the target readers of the significance of the conducted research, authors would scarcely use hedging devices. Therefore, I expect that the occurrence of passive voice or epistemic modals would be low. Furthermore, in abstracts, authors acquaint the reader with the research aim, the methodology, and the results in a succinct form. For this reason, I expect that there would be a high frequency of factive verbs (objective language in general) rather than non-factive verbs (evaluative language).

In discussions, scholars normally describe how they have carried out their study, what technique they used, or what approach they adopted. Then they evaluate the results in light of the adopted approach, report on the shortcomings and their possible impact on the final results, and express their stance on how successful this approach was in achieving the desired effect. For this reason, I expect that they would adopt a more personal approach and there would be a low frequency of passives.

In conclusions, scholars would make their claims about the examined phenomenon or the effectiveness of the adopted approach based on the previously discussed research results and would try to make predictions about the possible implications that the achieved results can have on the future development of the analyzed phenomenon.

Since authors tend to express some personal evaluation in both discussions and conclusions, there should be a high occurrence of epistemic modal auxiliary verbs, as well as non-factive/evaluative verbs denoting the degree of certainty with which authors make statements.

3.2.4 Cohesion

Cohesion is a semantic concept which deals with the relations of meaning in a text, and it is one measure of textual coherence. When the interpretation of one entity in a text is dependent on or recoverable from that of another entity that is also present in the text, it is said that these entities are in a cohesive relation and form a cohesive tie (Halliday & Hassan, 1976). Such relations can be signaled grammatically, for example, by means of references (coreferences), substitutions, and ellipses. The examples in Table 3 illustrate cohesive relations marked by grammatical cohesive devices identified manually in the section types of a CL research article (Source File: D10-1065-parscit.130908.xml).

Table 3 Examples of grammatical cohesive relations and the section type in which they have been manually identified.

Cohesive device	**Example**	**Section type**
Reference	Suppose on a common data set, **the sets of alignment links produced by two aligners** are A and B, we compute **their** agreement as [...]	Discussion
	This work empirically studies the performance of **these two classes of alignment algorithms** and explores strategies to combine **them** to improve overall system performance.	Abstract
Substitution	The difference from that work is that **our focus** is to leverage complementary alignment algorithms, while **theirs** was to leverage alignments of different lexical units produced by the same aligner. (Nominal Substitution)	Related Work

Cohesive relations can also be marked lexically by means of lexical repetitions, collocation chains (i.e., by words that tend to co-occur), or by various semantic relations which indicate that the entities have similar meanings (synonyms) or opposite meanings (antonyms), form a part-whole relation (meronyms), or denote a supertype-subtype relation (hypernym-hyponym) (Halliday & Hassan, 1976).

Table 4 Examples of lexical cohesive devices and the section type(s) in which they appear.

Cohesive device	**Example**	**Section Type**
Lexical repetitions	**data, training, algorithms, word alignments, improvements, methods, we, etc.**	Across the section types
Collocation chains (co-occurring)	**statistical machine translation (MT) - supervised/unsupervised methods - algorithms - training data - MT performance, etc.**	Across the section types
Synonyms (similar)	**experimental - empirical, heuristic**	Introduction
Antonyms (opposite)	**modest** v. **significant** improvements **theoretical v. empirical**; **coarse v. fine-grained** alignments	Abstract Introduction

	similar v. different method **long v. short** sentences	Related Work Discussion
Meronym (part-whole)	**MT system – pipeline**	Introduction
Hypernym-hyponym (general-specific)	**languages - English, Chinese, Arabic** **genres - newswire, weblog**	Abstract

Another means of marking cohesive relations is through the use of connectives. These can take the form of subordinating conjunctions such as *since, because, before, etc.;* coordinating conjunctions such as *and, or*, discourse adverbials like *moreover, therefore,* and PPs such as *on the one hand/on the other hand, by contrast, in the meantime, etc.*, which can establish *expansion, contingency, comparison/contrast, temporal* discourse relations between spans of text called the arguments of the connective (Webber et al., 2007). Furkó (2020) emphasizes the importance of connectives in discourse analysis by pointing out that they facilitate the understanding of these relations by guiding the readers to the author's intended interpretation of the existing connections and ruling out the unintended ones. Depending on whether the connectives are overtly expressed or inferred from the context, they can be of two types: explicit and implicit respectively. McNamara et al. (2009) also point out that connectives along with coreferences are central linguistic indices of cohesion that facilitate text comprehension.

Table 5 shows examples of both types, as well as the type of discourse relation each of them establishes between the two successive utterances. The types of discourse relations are based on the classification scheme used for the annotation of the sense relations that hold between the arguments of the connectives in the Penn Discourse Treebank Corpus (PDTB). The first argument of the connective is italicized and the second argument is underlined.

Table 5 Examples of explicit and implicit connectives, the type of discourse relation they establish between the argument sentences, and the section types in which they have been manually identified.

Connective	**Example**	**Discourse Relation**	**Section Type**
Explicit	*A main focus of much previous work on word alignments is on theoretical aspects of the proposed algorithms.* **In contrast,** the nature of this work is purely empirical. *Some studies leveraged other types of differences between systems to improve MT.* **For example**, de Gispert et al. (2009) combined systems trained with different tokenizations. Source: D10-1065-parscit.130908.xml	Comparison:: Contrast/ Juxtaposition Expansion:: Exemplification	Introduction Related Work
Implicit	*Modest improvements were achieved by taking the union of the translation grammars extracted from different alignments.* **IMPLICIT = In contrast,** Significant improvements (around 1.0 in BLEU) were achieved by combining outputs of different systems trained with different alignments. Source: D10-1065-parscit.130908.xml	Comparison:: Contrast Juxtaposition (a suggested implicit relation)	Abstract

The manual identification of such examples is intended to demonstrate that the sections of CL research articles can be a rich resource of various types of cohesive devices. The present study, however, focuses only on three of these devices -- coreferences, lexical repetitions, and explicit connectives. By exploring their distribution across the section types, I seek to ascertain whether the dominance of a particular cohesive marker can be motivated by the purpose of the section type. The hypothesis is that the section types will differ in terms of the frequency of occurrence of the different cohesive devices and that each section will have a predominant cohesion marker that contributes to its overall coherence. The frequency/dominance of the cohesive device can be explained by the hypothesized section intentions. The analysis of the distribution of the rest of the cohesive markers is left for future work.

Previous studies, among others Witte & Faigley (1981) and McNamara et al. (2009), acknowledge that cohesion is an important measure of text quality and text readability.
By analyzing academic students' essays, Witte & Faigley (1981) found that the use of cohesive devices and the frequency of cohesive ties can reflect students' writing proficiency and creative skills. Cohesive devices can also provide insights into the domains of reference present in the text, how they are retained throughout it, what are their realizations, and how the authors structure their argumentation. However, they point out that the high frequency of cohesion alone does not necessarily improve text readability and overall coherence. In fact, a text can be highly cohesive, which means that there is a high frequency of overtly expressed cohesive markers or repeated lexical items, which indicate ties, but they may not establish relations between adjacent or nonadjacent utterances simply because each utterance answers a different question, and does not contribute to the recognition of the overall discourse intentions. An example that illustrates very well this discrepancy between the presence of cohesive devices and the absence of overall coherence is the following provided by Enkvist (1990):

(2) *My car is black. Black English was a controversial subject in the seventies. At seventy most people have retired. To re-tire means to put new tires on a vehicle. Some vehicles such as hovercraft have no wheels. Wheels go round.*

The example above clearly shows that despite the presence of lexical repetitions, which seemingly establish relations between the adjacent utterances, the utterances altogether do not form a unified whole since each of them achieves a different communicative goal. This observation is also in line with Witte and Faigley's proposition that surface cohesive markers are only one factor that contributes to overall coherence and text quality. However, other factors such as the author's audience design and the intentions of the text can come into play when it comes to text readability and the understanding of the existing relations among the entities in the text.

The scholars went further to explain that the selection and the frequency of use of cohesive devices can be influenced by the authors' perception of their audience's expertise in the analyzed subject matter. The results from their observations show that the higher the competence of the addressee in the discussed topic, the lower the frequency of use of explicit cohesive markers (Witte & Faigley, 1981). This can be explained by the fact that for experts and researchers experienced in a particular field, it is easier to retrieve/identify existing discourse relations such as *cause-effect, elaboration, contrast, justification,* etc., or relations between entities without these being overtly signaled in the text. These observations suggest that shared knowledge (common ground) between the authors and their addressees can play a significant role in motivating the author to adjust his expression/production of utterances to their recipients. Vanlangendonck et al. (2013), for example, examined how the shared knowledge (i.e., the information available to both the speaker and the addressee) and the privileged knowledge (i.e., the information available only to the speaker) can influence this adjustment process called audience design. They investigated how the shared knowledge (the common ground) and the privileged knowledge (the privileged ground) affect the production of referring expressions, in particular. The results from their experiment showed that the speakers' utterance planning is partially constrained by both the shared and the privileged knowledge, i.e., they tend to adapt their expression by taking into account the addressees' perspective. More specifically, one of the observations suggested that when speakers have access to both shared and privileged knowledge during the stage of utterance planning, speakers tend to ignore the privileged information and enforce the shared knowledge by using referring expressions in order to ensure the success of the communicative situation. Their research results confirmed previous views according to which if speakers fail to consider the common ground and fail to ignore the privileged ground, this can result in addressees' confusion and ineffective communication. Another factor that can improve the overall text processing and understanding is the text purpose. The present study, however, does not focus on the effect of text intentions on text readability but rather examines the effect of the text type on the choice and the frequency of occurrence of cohesive devices across the section types.

Cohesion has been analyzed by retrieving instances of coreferential relations, relations marked by lexical repetitions, and explicit connectives.

3.2.4.1 Coreference

Coreference is a type of referential relation in which the entities that form a tie -- the anaphor and its antecedent -- have the same referent in the real world. The anaphor is typically realized by proforms, repeated NPs, or modified NPs. The coreferential relations have been retrieved with the StanfordCoreNLP Module, which displays them in the form of coreference chains. Each coreference chain stands for an anaphor-antecedent relation and can contain two or more mentions of the same entity. Coreference chains can give us insights into which referents (types of information) are persistent throughout the text and what are their realizations.

The frequency of occurrence of coreference relations has been measured by extracting the total number of coreference chains and normalizing it by the total number of sentences.

3.2.4.2 Lexical Repetitions

Relations among entities in a text can be marked not only by words/phrases pointing backward or forward to other entities in the text but also by lexical repetitions. The role of lexical repetition in text production has been recognized and analyzed frequently, among others, Halliday & Hassan, 1976; Hoey, 1991. According to them, repetition, as a marker of lexical cohesion, is not only a crucial mechanism that improves overall *discourse texture*, i.e., the potential of a sequence of utterances to behave as a unit, but it is also a key writing strategy of the authors to signal the continuity/retention of the topic(s) they have dealt with throughout the text. On the one hand, this can ease text processing, and on the other, it can demonstrate the authors' skills in building a clear text structure, in which there are no information gaps and in which the authors attempt to present their arguments consistently and logically.

In general, lexical repetitions tend to be frowned upon in creative writing since they are treated as the ultimate indicators of the writers' limited vocabulary or absence of writing proficiency. Authors are, therefore, encouraged to use fewer repetitions and to employ synonyms instead or other types of words whose meaning is close or similar to that of the target word. However, Adorján (2013) relevantly reports that the necessity of lexical repetitions in texts can be genre-motivated. In scholarly texts, for example, the repetitions of subject-specific concepts/terms cannot be replaced by synonyms since this might not only hinder the readability of the text but can also result in the readers misinterpreting the authors' intentions. As has already been demonstrated with coreference relations, authors tend to repeat not only field-specific words/phrases but also proper nouns, i.e., names of peer scholars, especially when they compare their approach/methods with those adopted by the mentioned scholars. Authors may also repeat personal pronouns, especially self-referring pronouns like *we* when they want to focus readers' attention on their research objectives, on their experimental design, on their findings, and finally on their possible research contribution. Therefore, lexical repetitions can be considered a critical rhetorical and organizational tool employed to ensure that each piece of information appearing in the text is linked to other pieces and that there are no information gaps between the parts of an utterance. One of the theories that offer a model for the analysis of coherent structures -- the Rhetorical Structure Theory (*RST*) -- suggests that what makes a sequence of utterances coherent is "the absence of non-sequiturs or gaps", i.e., each piece of information has a particular function and that there should be a plausible motivation/evidence for its presence (Mann & Thomson, 1988). A surface-based mechanism, which can signal this absence of gaps as the text unfolds, is the repetition of lexemes.

Recent studies on cohesion and coherence, e.g., (Wang & Zhang, 2019), emphasize the crucial contribution of lexical cohesion to text coherence both on a local and a global level in research articles. In an experiment, they explored the distribution of the different types of lexical cohesive devices such as repetitions, synonyms, antonyms, meronyms, hyponyms in the Introduction-Method-Result-Discussion (*IMRD*) structure of 30 research articles in applied linguistics. They found that lexical repetition is by far the most frequently used lexical device across all subsections. To be more specific, 91% of all detected lexical cohesive devices were represented by repetitions, 3% by antonyms, 3% by synonyms, 2% by meronyms, and the hyponyms were the least represented -- only 1%. Wang and Zhang also pointed out that the frequency of occurrence of a particular cohesive device can be correlated with the function of the section. For example, lexical repetition was most frequent in introductions, results, and discussion/conclusion sections. Their explanation for this high frequency is that in introductions, authors try to activate readers' memory of keywords, in results -- to focus readers' attention on the outcomes, and in the discussion/conclusion section -- to repeat the objectives established in the introduction section and to comment on the extent to which these objectives were achieved in light of the outcomes presented in the results section. What is more, meronyms were most frequent in the introduction and methodology sections because authors tend to define the key terms. Introductions and discussion/conclusion sections demonstrated the highest frequency of synonyms. However, what Wang and Zhang noticed is that the synonym pairs were of a different nature in the two sections in the sense that readers would be able to understand the synonym relations in introductions without preliminary knowledge, whereas in the discussion/conclusion section, the synonym pairs were very much topic-specific or the meaning of the synonym pairs was constrained by the nature of the analyzed topic.

When it comes to the antonyms and the hyponyms, they appeared mostly in introductions and their meanings were again topic-dependent.

The research presented here can be viewed as a natural continuation and an addition to Wang and Zhang's study results on the frequency of lexical repetition in research articles. The present approach attempts not only to assess the distribution of repetition across the constituent sections but also to elaborate on the effect of the section purpose on the frequency of repetitions. It also investigates how lexical repetition contributes to text coherence on a local level (inside each constituent section) and on a global level (between the sections) by analyzing the topics that are retained throughout the article and the topics which are activated only in particular sections (i.e., section-specific topics).

During the first stage of my work, I analyzed the frequency of lexical repetitions across the section types. What counts as a lexical repetition is a content word (noun, verb, adjective, adverb), which appears 2 or more times in a text, thereby forming a lexical chain. With the help of Python and its library for NLP SpaCy 2.2.4, the lemmas of the content words for each text were collected in a list, then the duplicate lemmas were identified and stored in a dictionary with their respective frequencies. Details on the repeated words and their counts per section type are available on GitHub and on GUDe in the directory **Linguistic Mechanisms > supplements > lexical chains**.

The example below illustrates the measurement of the frequency of repeated words across section types.

(3) Source File: J12-3005-parscit.130908.xml file

__Adjectives__ are one of the most elusive parts of speech with respect to meaning. For example, it is very difficult to establish a broad __classification__ of __adjectives__ into __semantic classes__, analogous to a broad ontological __classification__ of nouns.

The following dictionary, extracted from the ***introduction*** section of the source file *J12-3005-parscit.130908.xml*, contains a sequence of repeated *word:count* pairs. The counts of all repeated words were summed up and normalized by the total number of tokens in the introduction section of this file.

{**'adjective': 20,** 'part': 2, 'meaning': 3, **'example': 9,** 'very': 2, 'establish': 2, 'broad': 2, **'classification': 7**, **'semantic': 12**, **'class': 14**, 'noun': 4, 'nirenburg': 2, 'language': 2, 'mail': 3, 'e': 2, 'submission': 2, 'receive': 3, 'work': 4, **'article': 4**, 'first': 4, '**computational': 5**, 'linguistic': 4, **'task': 4,** 'give': 3, 'property': 4, 'other': 3, 'acquisition': 5, 'study': 2, 'be': 2, 'exception': 2, 'can': 3, 'will': 2, 'test': 2, 'different': 5, **'empirical': 3**, **'problem': 4**, **'polysemy': 10**, 'fact': 2, **'sense': 21**, 'such': 4, 'exhibit': 2, 'similar': 2, 'alternation': 7, 'regular': 8, 'systematic': 2, 'briscoe': 2, 'research': 3, 'present': 3, 'therefore': 3, 'model': 7, 'instance': 2, 'derive': 4, 'economy': 4, 'translate': 3, 'cheap': 3, 'see': 2, 'correspond': 2, 'recovery': 2, 'economysuffix': 2, 'trouser': 2, 'familiar': 5, 'family': 3, 'love': 3, 'show': 2, 'meeting': 2, 'face': 2, 'boy': 2, 'lovely': 2, 'relationship': 2, 'goal': 2, 'belong': 4, 'individual': 2, 'related': 2, 'human': 3, 'theoretical': 2, **'approach': 3**, 'direction': 2, 'e.g.': 2, 'ai': 2, 'more': 2, **'section': 5**}

In the example above, the sum of all counts of repeated words, which equals *316* is divided by the total number of tokens -- *1008*, which results in a frequency score of *0.31349206*.

The example shows that the most repeated words, which also form the longest lexical chains, are the nouns *sense, adjective, semantic, classification, polysemy, class,* etc. The semantics of these words allows us to make predictions about the topic under discussion, which in this case, most probably is related to the semantic classification of adjectives for a particular language/polysemous adjectives and their classification with respect to their senses. This shows that the introduction section is laden with context-bound/topic-dependent lexical chains.
Considering the purposes/functions of the individual sections and Wang and Zhang’s findings, the initial expectations are that introductions and discussions/conclusions would contain the highest frequency of lexical repetitions since, in introductions, authors tend to **acquaint** readers with the major research problem that they would like to explore and find solutions to, and also to **provide** a brief overview of the organization of the article’s content. Therefore, one would expect a high frequency of repetitions of subject-specific/domain-specific words/phrases such as the ones in the example above (in the current research within the context of computational linguistics), or words referring to the organization of the article such as *section, paragraph,* etc. that facilitate the reading process.

In discussions, alternatively, in conclusions, authors usually comment on the research results and would relate these results to the aims/questions formulated in the introduction section and possibly juxtapose their results with the results of previous studies mentioned in related works. In other words, in discussions or conclusions, authors would be expected to reactivate some of the information presented in the previous sections. Context-bound words/phrases, as well as words referring to the results or the presentation of the results, may be repeated frequently.

3.2.4.3 Explicit Connectives

The analysis of explicit connectives and the types of discourse relations they can signal between pieces of information has attracted a lot of attention in recent years due to its possible implementations in the development of systems for the automatic recognition of discourse relations (Pitler & Nenkova, 2009), as well as for the automated evaluation of text argumentation and text coherence (McNamara & Graesser, 2011).

The distribution of the types of discourse relations across the constituent sections in the present work is measured by the occurrence of explicit connectives signaling *temporal, comparison, contingency,* or *expansion* relations inter- or intrasententially. These four relation types are based on the classification scheme used for the annotation of the sense relations that connectives mark between their arguments in the Penn Discourse Treebank Corpus (*PDTB*) (Weber et al., 2007). The analysis of the distribution of the different types of relations can help us understand how authors construct, develop, and support their arguments in the course of the article writing. It can also shed light on whether and how the individual sections impose constraints on the frequency of explicit connectives and the occurrence and distribution of a particular type of discourse relation.

Connectives such as *previously, simultaneously, thereafter,* etc. indicate if two or more events (actions or states) are in a synchronous relation, i.e., are taking place simultaneously, or are in an asynchronous relation, i.e., one of the events precedes or follows the other. The hypothesis is that introductions or related works would contain a high frequency of connectives marking temporal relations since authors make references to previous studies and make comparisons between past and present procedures.

Moreover, connectives such as *although, however, nevertheless,* etc. signal opposing/contrastive pieces of information. Since authors typically compare their current results with their initial hypotheses or explain to what extent the outcomes from the conducted experiment/tested model confirm or deviate from their expectations, I expect that discussions would contain a high frequency of such connectives.

Consequently, therefore, hence, because, if, etc. mark a cause-effect, reason, result, or conditional relation. Such relations can be prevalent in introductions because scholars tend to give motivation for the approach/research procedure they have adopted in order to provide a solution to the research problem. In discussions, there might also be a high frequency of such connectives since authors explain in detail what they have done, how they have done it, and to what results their actions have led.

Finally, connectives such as *moreover, besides, firstly/secondly…,* etc. signal addition/elaboration, exemplification, or listing of a sequence of items. The expectations are that all section types would demonstrate a high frequency of them.
The full list of explicit connectives is available in Appendix 3.

Chapter 4 Evaluation of the Research Results

The current section provides a quantitative assessment of the distribution of the analyzed features across the five section types and offers an interpretation of the research results.

The dataset consists of five dependent samples (sections). They are dependent in the sense that each row of the dataset corresponds to one research article and the aim is to verify if the frequency of occurrence of a linguistic feature varies across these five samples under the effect of the section type. The difference in means was tested for statistical significance for the different features. Since the assumptions of normality and equal variances of the data across the section types were violated, the non-parametric Wilcoxon signed-rank test (Rey & Neuhauser, 2011) was conducted (the statistical details for each feature are available on GitHub, as well as on GUDe in the directory **Linguistic Mechanisms > statistical_analysis**). If the result of the test showed overall significance, a pairwise test with a Bonferroni correction was conducted to determine between which section pairs, in particular, the differences in means are statistically significant.

4.1 Nominalization

4.1.1 NP Count

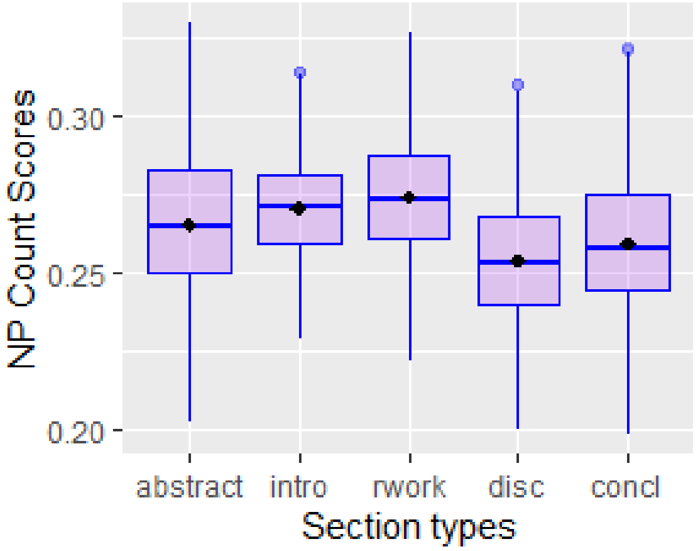


Figure 3 Distribution of NPs across the section types: abstract, intro (introduction), rwork (related works), disc (discussions), concl (conclusions). The blue horizontal line represents the median and the black point represents the mean. In the figure above, related works demonstrate the highest frequency of occurrence of NPs, whereas discussions have the lowest frequency.

The results show that related works (0.275) and introductions (0.272) demonstrate the highest frequency of NPs, followed by abstracts (0.266), conclusions (0.260), and discussions (0.255). The result from the Wilcoxon signed-rank test confirms that the differences in means are statistically significant with **p-value < 2.2e-16, V = 15654810**.

The pairwise comparisons also confirm that the differences in means between all sections are statistically significant. The statistical details from the pairwise comparisons are given in Table 6.

Table 6 NP Count. Results from the pairwise comparisons.

	abstract	introduction	related work	discussion
introduction	4.4e-10	-	-	-
related work	< 2e-16[3]	5.7e-05	-	-
discussion	< 2e-16	< 2e-16	< 2e-16	-
conclusion	1.3e-10	< 2e-16	< 2e-16	1.9e-09

A possible explanation for the high frequency of NPs in related works and introductions is that authors cite previous studies in these sections to set the scope of their own study, so they tend to make use of domain-specific notions or proper nouns when they cite the authors of the conducted studies.

4.1.2 NP Length

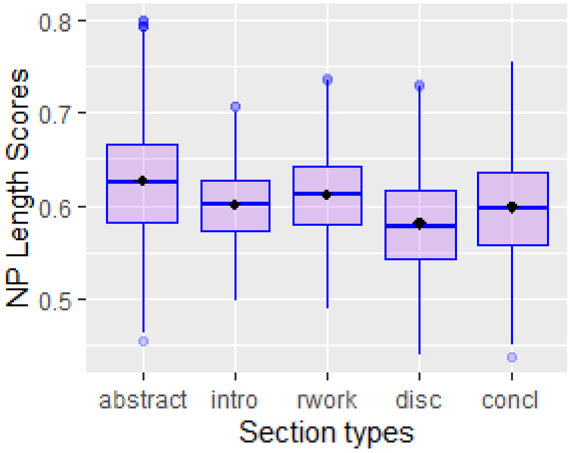


Figure 4 Distribution of NP Length. Abstracts have the longest, whereas discussions have the shortest token-based NPs.

The graph displays the variation in the mean token-based NP Length across the five section types. The result from the Wilcoxon signed-rank test confirms that the differences in means are statistically significant with **p-value < 2.2e-16, V = 15654810**. The pairwise comparisons also show that the null hypothesis can be rejected for all pairs, except for the introduction-conclusion pair where the difference in means is not statistically significant.

[3] 2.2e-16 is the lowest value that R prints out. It is a scientific notation for a number very close to zero.
< 2.2e-16 indicates that the difference between the compared groups is very large and is statistically significant (smaller than the threshold 0.05).

Table 7 NP Length. Results from the pairwise comparisons.

	abstract	introduction	related work	discussion
introduction	< 2e-16	-	-	-
related work	5.5e-08	4.3e-07	-	-
discussion	< 2e-16	< 2e-16	< 2e-16	-
conclusion	< 2e-16	0.99	5.7e-08	5.8e-11

Although related works and introductions demonstrate the highest frequency of NPs, as has been pointed out previously, it seems that abstracts contain the longest token-wise NPs (0.629), followed by related works (0.613), introductions (0.603), conclusions (0.602), and discussions (0.583).

The differences in the NP Length across the sections can be correlated with the purpose of the section and the size constraints -- the shorter the text, the longer the NP. These two factors encourage writers to adapt their expression to the section-specific requirements by making use of information-burdened noun phrases in abstracts. In discussions, by contrast, the smaller length suggests that authors tend to adopt less frequently lengthy NPs since the information they convey can be encoded in the form of VPs rather than condensed in an NP structure. These findings support the initial expectations and confirm the claims in Ivanova (2020). The types of NPs that appear in discussions and conclusions are personal pronouns with which authors would normally focus readers' attention on the set of activities that were carried out during the research procedure, the research results, and their possible interpretation

4.1.3 NP Complexity

4.1.3.1 Adjective (Adj) Dependents

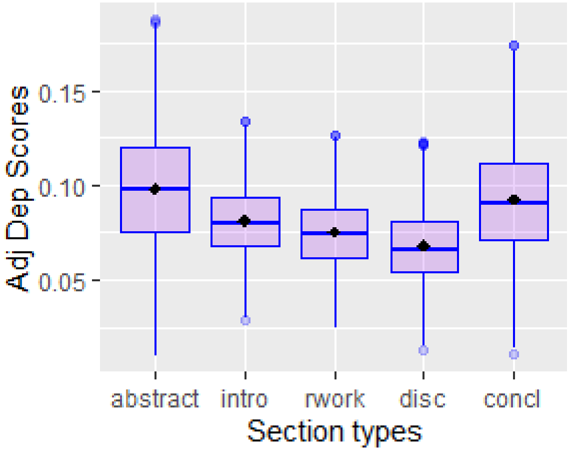


Figure 5 Distribution of NPs containing Adjective Dependents. Abstracts demonstrate the highest frequency of NPs containing Adjectives, discussions the lowest.

The graph shows the distribution of NPs in which the noun head takes adjective dependents (e.g., *an <u>efficient</u> <u>statistical</u>* ***ranking***, *a <u>domain-specific</u> <u>semantic</u>* ***representation***[4], etc.). The results from the statistical analysis show that abstracts have the highest occurrence of NPs containing adjectives with a score of 0.10, followed by conclusions with a score of 0.094. In abstracts and conclusions, the size restrictions of abstracts force scholars to express content very economically. This means that one should find more complex NPs in abstracts and conclusions than in other parts of a paper and this is signaled by heavy modification inside the NP structure.

The rest of the section types demonstrate a lower occurrence of such syntactically complex NPs -- introductions (0.082), related works (0.077), and discussions (0.070). The data in all samples were normally distributed and the result from the Wilcoxon signed-rank test confirms that the differences in means are statistically significant with **p-value < 2.2e-16, V = 15621255**. The pairwise comparisons also show that the differences in means are statistically significant between all pairs.

[4] The adjective dependents of the noun head are underlined; the noun head appears in bold.

Table 8 Adjective Dependents. Results from the pairwise comparisons.

	abstract	introduction	related work	discussion
introduction	< 2e-16	-	-	-
related work	< 2e-16	7.1e-10	-	-
discussion	< 2e-16	< 2e-16	2.4e-16	-
conclusion	0.00048	< 2e-16	< 2e-16	< 2e-16

4.1.3.2 Noun Dependents (Noun)

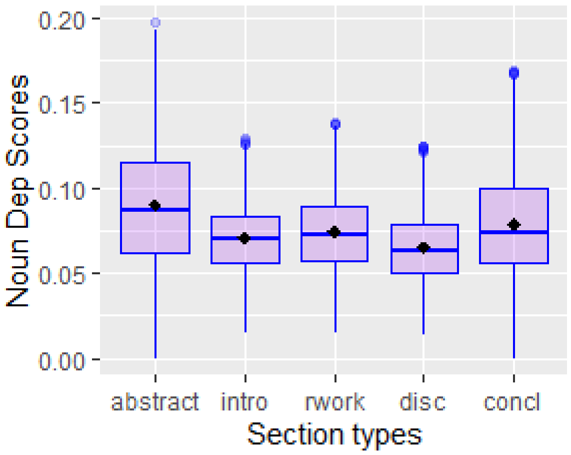


Figure 6 Distribution of NPs containing Noun Dependents. Abstracts demonstrate the highest frequency of such NPs, discussions the lowest.

Section types also differ in terms of the frequency of occurrence of NPs containing noun dependents, examples of which are *a spoken* _dialogue_ _language_ ***system***[5], *natural* _language_ _generation_ ***systems***, *Many corpus-based* _machine_ _translation_ ***systems***, *a spoken* _dialogue_ _language_ ***system*** *for making air travel plans over the telephone, etc*. The distribution of NPs containing noun dependents also confirms that abstracts (0.091) have the most syntactically complex NP structures, again followed by conclusions (0.081), related works (0.075), introductions (0.073), and discussions (0.068). The result from the Wilcoxon signed-rank test confirms that the differences in means are statistically significant with **p-value < 2.2e-16, V = 15610078**. The differences in means between all pairs turn out to be statistically significant.

[5] The noun dependents of the noun head are underlined; the noun head appears in bold.

Table 9 Noun Dependents. Results from the pairwise comparisons.

	abstract	introduction	related work	discussion
introduction	< 2e-16	-	-	-
related work	< 2e-16	0.00064	-	-
discussion	< 2e-16	9.2e-12	< 2e-16	-
conclusion	< 2e-16	1.5e-12	0.00015	< 2e-16

4.1.3.3 Prepositional Phrase (PP) Dependents

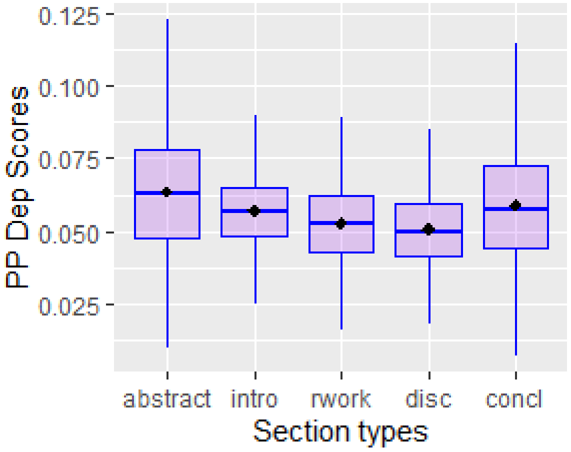


Figure 7 Distribution of NPs containing PP Dependents. Abstracts demonstrate the highest frequency of NPs containing PP dependents, discussions the lowest.

The section types differ in terms of the distribution of NPs in which the noun head takes a PP dependent (e.g. *an empirical* ***evaluation*** *of an adaptive mixed initiative spoken dialogue system*[6]*, their* ***strategies*** *for preventing, identifying and repairing problems,* etc.*)*. Abstracts demonstrate the highest frequency of PP dependents (0.064), followed by conclusions (0.60), introductions (0.58), related works (0.53), and discussions (0.51). The result from the Wilcoxon signed-rank test confirms that the differences in means are statistically significant with **p-value < 2.2e-16, V = 15615666**. The pairwise comparison shows that the differences in means are statistically significant between all pairs, except for that between introductions and conclusions.

[6] The PP dependent of the noun head is underlined; the noun head appears in bold.

Table 10 PP Dependents. Results from the pairwise comparisons.

	abstract	introduction	related work	discussion
introduction	2.1e-15	-	-	-
related work	< 2e-16	< 2e-16	-	-
discussion	< 2e-16	< 2e-16	0.0031	-
conclusion	1.8e-06	0.1642	< 2e-16	< 2e-16

4.1.3.4 Past Participle (VBN) Dependents

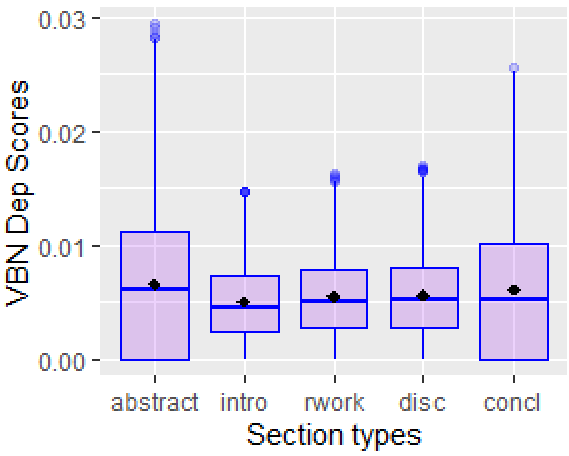


Figure 8 Distribution of NPs containing Past Participle (VBN) Dependents. Abstracts and conclusions demonstrate the highest frequencies of NPs containing Past Participles as dependents, introductions have the lowest frequency.

The graph shows the distribution of NPs whose heads take past participle dependents (VBNs), examples of which are *a* ***classifier*** *trained with only automatic features*[7], ***information*** *encoded in the top level nodes*, etc. Abstracts demonstrate again the highest frequency (0.007), followed by conclusions (0.006), discussions (0.006), related works (0.006), and introductions (0.005). The result from the Wilcoxon test confirms that the null hypothesis can be rejected with **p-value < 2.2e-16, V = 9169903**. The pairwise comparisons show that the differences in means are statistically significant between the following pairs: abstract-

[7] The VBN dependent of the noun head is underlined; the noun head appears in bold.

introduction, abstract-related work, abstract-discussion, introduction-related work, introduction-discussion, and introduction-conclusion.

Table 11 Past Participle Dependents. Results from the pairwise comparisons.

	abstract	introduction	related work	discussion
introduction	1.1e-07	-	-	-
related work	0.0175	0.0055	-	-
discussion	0.0218	0.0010	1.0000	-
conclusion	0.1358	0.0044	1.0000	1.0000

In light of the quantitative assessment of the NP Complexity across the five section types, it can be concluded that abstracts demonstrate the highest frequency of syntactically complex NPs. These results are also in line with the NP Length feature and confirm the hypothesis that in abstracts, authors seek to familiarize their audience with the article's central topics in a space-efficient manner, so they tend to use complex NPs in which nouns take dependents of various types and numbers. Conclusions also demonstrate a high frequency of complex NPs which again can be motivated by the size constraints, on the one hand, and by the section purpose, on the other hand. In conclusions, authors tend to refer back to the content presented in the previous sections in order to make claims about the extent to which the research aim was achieved with regard to the adopted approach/methods. This encourages them to remind the reader of the central topics activated already in the abstract such as the aim, the methods, the results, and the possible contribution of the study.

4.2 Self-Mentions

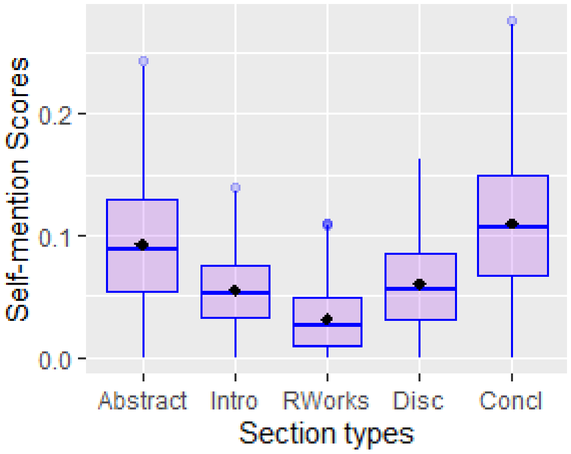


Figure 9 Distribution of Self-Mentions. Conclusions demonstrate the highest frequency of self-mentions, related works demonstrate the lowest frequency.

The graph shows that the section types also differ in terms of the distribution of self-mentions: abstracts (0.094), introductions (0.056), related works (0.033), discussions (0.062), and conclusions (0.113).

The result from the significance test shows that the null hypothesis can be rejected with **p-value < 2.2e-16, V = 13150756**. The results from the pairwise comparisons also confirm that the differences in means are statistically significant between all pairs.

Table 12 Self-Mentions. Results from the pairwise comparisons.

	abstract	introduction	related work	discussion
introduction	<2e-16	-	-	-
related work	<2e-16	<2e-16	-	-
discussion	<2e-16	0.0025	<2e-16	-
conclusion	<2e-16	<2e-16	<2e-16	<2e-16

As my hypothesis predicted, authors tend to make references to themselves mostly in conclusions and abstracts. This can be explained by the fact that in conclusions, authors normally make their claims about the investigated phenomenon, and present these claims in the form of personal evaluations since their interpretations are limited to the scope of the presented research results. Moreover, in conclusions, authors can express their beliefs/expectations on how the research community can benefit from their results or what possible implementations these results can find in future studies/experiments, or how they can be used to improve the functionality of a system. Therefore, in conclusions, self-mentions can have a focusing function -- to draw the readers' attention to some content of an evaluative nature, i.e., authors' personal evaluation of the achieved results or their interpretations/expectations. In abstracts, self-mentions can also possess a focusing function; this time, to focus readers not on evaluations but rather on facts/objective content such as the aim, the methodology, and the results. These interpretations can be supported by the frequency of self-mentions co-occurring either with factive (non-evaluative) verbs or with non-factive (evaluative) verbs shown in Figure 10. Indeed, in conclusions, self-mentions tend to co-occur more frequently with non-factive (evaluative) verbs and abstracts -- more frequently with factive (non-evaluative) verbs.

Examples of self-mentions co-occurring with non-factive (evaluative) verbs in conclusions:

(4) Source File: A94-1006-parscit.130908.xml

As the need for efficient knowledge acquisition tools becomes widely recognized, ***we hope that*** *this experience with termight will be found useful for other text-related systems as well.*

(5) Source File: D08-1057-parscit.130908.xml

We *also* ***showed that*** *domain-specific patterns, schematic word-pair co-occurrences in this case, can be acquired from a limited amount of data as indicated by modest performance gains for content selection using schemata information.* ***We postulate that*** *this is particularly true when dealing with homogeneous data.*

(6) Source File: D09-1012-parscit.130908.xml

We believe that *sharing these fragments with the NLP community and studying them in more depth will be useful to identify new, relevant features for the characterization of several learning problems.*

Examples of self-mentions co-occurring with factive (non-evaluative) verbs in abstracts:

(7) Source File: W96-0101-parscit.130908.xml

We demonstrate that*, besides providing good estimates for disambiguation, word classes solve some of the problems caused by sparse training data.*

(8) Source: W10-1819-parscit.130908.xml

***We establish that** the PropBank scheme is applicable to clinical Finnish as well as compatible with the SD scheme, with an overwhelming proportion of arguments being governed by the verb.*

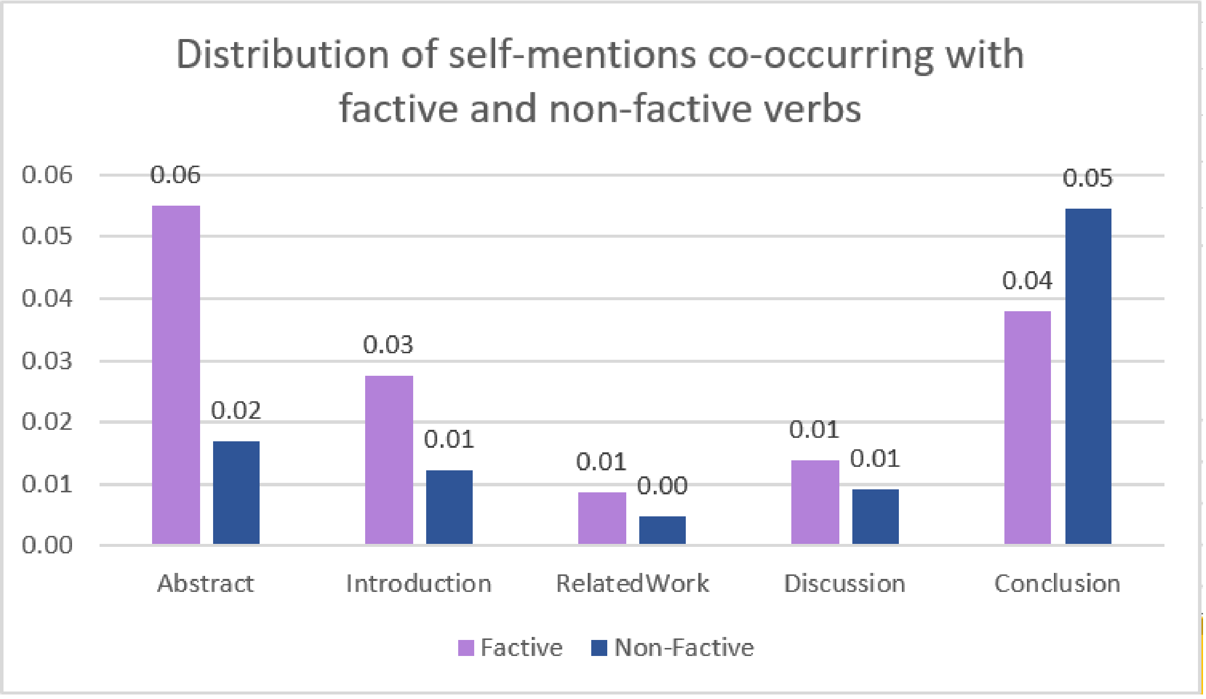


Figure 10 Distribution of self-mentions co-occurring with factive and non-factive verbs.

Contrary to my initial expectations, discussions do not demonstrate a high frequency of self-references. As explained, in discussions, authors tend to reactivate previous knowledge of the series of actions that were conducted or the strategies they followed to achieve the reported results, so they might do so by using impersonal constructions instead of self-references since the role of the authors can be inferred from context.

Finally, related works demonstrate the lowest frequency of self-mentions. This can be motivated again by the purpose of the section -- to link the current study to already conducted studies by presenting as objectively as possible the claims and the contributions of these studies. Therefore, in related works, authors would make references to peer scholars and their works more frequently than to themselves and their own works.

4.3 Hedging

4.3.1 Passive Voice

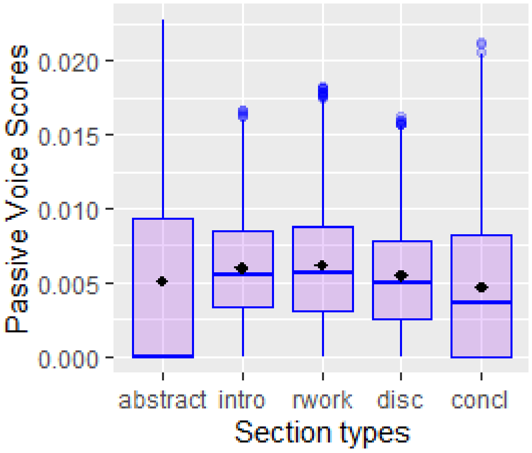


Figure 11 Distribution of Passive Voice. Related works demonstrate the highest frequency of passive voice, conclusions demonstrate the lowest frequency. In abstracts, the median is 0 and the mean is greater than 0 because more than half of the values in the abstract sample are 0 and the rest are positive values.

The figure shows the distribution of passives across section types. The total number of passive voice occurrences in a section was normalized by the total number of tokens.

The results show that the highest occurrence of passive voice is concentrated in related works (0.0066), and introductions (0.0063), followed by abstracts (0.0060), discussions (0.0059), and conclusions (0.0052). The test result shows that the null hypothesis can be rejected and there are samples whose differences in means are statistically significant with **p-value < 2.2e-16, V = 8692365**.

Table 13 shows concretely between which samples the differences in means are statistically significant: abstract-introduction, abstract-related work, introduction-discussion, introduction-conclusion, related work-discussion, related work-conclusion, and discussion-conclusion.

Table 13 Hedging marked by Passives. Results from the pairwise comparisons.

	abstract	introduction	related work	discussion
introduction	0.0027	-	-	-
related work	0.0011	1.0000	-	-
discussion	0.8922	0.0283	0.0047	-
conclusion	0.0632	1.5e-10	4.1e-12	1.0e-05

As expected, in related works, there is a high frequency of passives, which can be interpreted as a form of defence mechanism on the part of the writers. By using passives, authors seem to indicate that they simply report, as precisely as possible, processes and findings of previous studies, and constrain their personal liability to the assertions. This interpretation relates back to and confirms Hyland's (1998) writer-based function of hedging, according to which authors use impersonal constructions, in this case, passives, to avoid full commitment to the statements and observations made by other scholars.

Introductions also demonstrate a high occurrence of passives, which can be explained by the fact that in this section writers normally draw a general picture of the paper's content. They tend to put their research into context by making references to previous investigations of the subject matter. In the context of computational linguistics, there might be, for instance, references to previously tested language models or previous methods that have been adopted to improve the performance of particular models. Authors might comment on their peer scholars' observations in order to emphasize in what ways their current research can be considered an improvement over previous approaches, or in what ways their study contributes to the resolution of a processing problem or the improvement of a model's performance.

Moreover, the semantics of the most common past participle verbs that appear in passive voice constructions can be correlated, to a certain extent, with the intentions of the section and the jargon of computational linguistics. This becomes evident from the collection of past participles that are most representative of a section type. The computed values are based on **M**utual **I**nformation (**MI**) scores (the table with the scores is available on GitHub and on GUDe in the directory **Linguistic Mechanisms > MI scores > Past Participles (VBNs)**).

For example, the VBN *questioned* is a verb with one of the highest scores and its meaning can be correlated with the assumption that in introductions, authors usually present a research problem -- a concept/a method/an approach/results that are to be questioned/verified in order to check their credibility or a model to be tested in order to evaluate its performance. Other verbs that are prevalent in introductions are **dynamic** verbs, which can refer to the actions that have been taken by other scholars or by the authors themselves for the research problem to be addressed or a particular model to be tested, e.g., *predefined, selftrained, traced, accented, factorized, etc.*

It can be the case that in introductions, authors mention previous studies in order to contrast them with the current one by specifying the existing differences and by emphasizing the advantages of the current study over the previous ones. Related works can be regarded as natural continuations of introductions in which authors explore in greater detail the observations their peer scholars have made, the hypotheses they have formulated, and the results they have achieved.

In comparison to the passive uses in introductions and related works, passives in discussions seem to have a different function. In discussions, authors tend to revisit the activities/approaches/techniques they have used to conduct their research. Therefore, they would use passives not to evade commitment to the presented content, but rather to focus the recipients on the performed activities rather than on the agents, which are the authors themselves. The examples below from discussions illustrate this function of passives:

(9) Source File: D09-1096-parscit.130908.xml
The results are calculated *using 10-fold cross-validation.*
Accuracy is shown *for three tasks — nine-, three- and two-zone classification — using both line and zone-fragment classification.* ***Performance is compared*** *against a majority class baseline in each case.*

Passives in discussions can also be used for organization purposes -- to focus readers' attention on the tables or figures showing the research results. For example,

(10) Source File: D08-1049-parscit.130908.xml
Results for the two tasks ***are given*** *in Tables 4 and 5 and in Figures 1 and 2.*

The retrieved passive voice constructions for each file/article per section type are available on GitHub and on GUDe in the directory **Linguistic Mechanisms > supplements > passives.**

Depending on how passives are used in the sections, I claim that they can have both a hedging and a non-hedging function. Thus, if authors employ them to report on the observations that other scholars have made, as observed in introductions and related works, they detach themselves from the presented content, and therefore, such passives can be treated as hedges. Otherwise, when passives are used simply to list activities performed by the authors themselves or when the agents are predictable, as observed in discussions, they would no longer have the hedging effect.

Here are two examples containing passives found in the related work and the discussion sections to illustrate the proposed difference:

(11) Source File: A00-1012-parscit.130908.xml (Related work)
*It **was found** that the Penn TreeBank sentences were 86% correct and the system output 66% correct.* (hedging)

(12) Source File: D08-1049-parscit.130908.xml (Discussion)
*Results for the two tasks **are given** in Tables 4 and 5 and in Figures 1 and 2.* (non-hedging)

In light of the reported results so far, it becomes clear that the distribution of passives across the section types is in line with the distribution of self-mentions. Those sections which demonstrate a higher frequency of self-mentions -- abstracts and conclusions -- tend to have a lower frequency of passives. Related work seems to be the section-type in which authors adopt an objective/fact-oriented/less evaluative expression, which becomes evident from the highest frequency of passives and the lowest frequency of self-mentions. By contrast, conclusions demonstrate the lowest passive scores and the highest frequency of self-mentions, which confirms the hypothesis that authors tend to adopt a more personal/stance-oriented approach in this section.

4.3.2 Epistemic Modality

Epistemic modals are the most common linguistic means of explicitly qualifying commitment to the truth value of a proposition. These modals can take the form of modal auxiliaries, modal lexical verbs, modal adjectives, or adverbs. In the current study, I analyzed the distribution of modal auxiliaries and modal lexical (non-factive) verbs to check in which section(s) they are prevalent and how this frequency can be motivated by the purpose of the section-type.

4.3.2.1 Modal Auxiliary Verbs

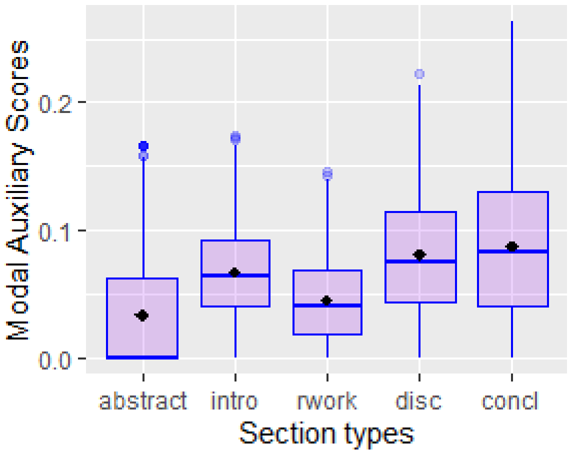


Figure 12 Distribution of Modal Auxiliary Verbs. Conclusions demonstrate the highest frequency of modal auxiliaries, abstracts demonstrate the lowest frequency. In abstracts, the median is 0 and the mean is greater than 0 because more than half of the values in the abstract sample are 0 and the rest are positive values.

Conclusions (0.089), discussions (0.084), and introductions (0.070) demonstrate higher occurrences of modal auxiliaries than related works (0.047) and abstracts (0.037). The results from the Wilcoxon test show that the null hypothesis can be rejected and it can be concluded that there is a significant difference between the compared means with **p-value < 2.2e-16, V = 9620691**. The results from the pairwise comparison also show that the difference in means is statistically significant between all pairs except for that between discussion and conclusion.

Table 14 Hedging marked by Modal Auxiliary Verbs. Results from the pairwise comparisons.

	abstract	introduction	related work	discussion
introduction	< 2e-16	-	-	-
related work	4.3e-09	< 2e-16	-	-
discussion	< 2e-16	6.5e-11	< 2e-16	-
conclusion	< 2e-16	6.5e-11	< 2e-16	0.34

The high frequency of modal auxiliaries in conclusions and discussions can be explained by the fact that in discussions/conclusions, authors tend to provide a personal interpretation of their study results. They also comment on how the potential limitations of the adopted approach might have affected the outcome. The use of modal auxiliaries, signaling writers' degree of certainty with which they make their statements, can be related to the accuracy-oriented function of hedges, according to which writers use them to indicate that they draw these inferences from the presented observations and since they acknowledge the existing limitations, they cautiously formulate their personal understanding of the research results. At the same time, they also leave room for the research community to make their own contribution by taking an active part in the reasoning and the understanding of the presented results, which correlates with the reader-oriented function of hedges.

Examples from the section types illustrating the use cases of modal auxiliaries in conclusions and discussions:

(13) Source File: W06-3001-parscit.130908.xml (Conclusion)
*[...] However, it **might** also be the case that in that kind of interactions no implicit referring expressions are used beyond the segmental level, because there is no such an extended context. [...]*

(14) Source File: A00-2037-parscit.130908.xml (Discussion)
[...] *In particular, we cannot conclude from the current study's small sample how strong the preference for using acknowledgment **might** be, especially among male subjects.* [...]

Moreover, in conclusions, scholars tend to make claims about the effectiveness of their approach/research procedure in achieving the desired effect in light of the discussed outcomes or comment on the possible contribution of the findings to future studies or their possible usability for the development of applications.

(15) Source File: W06-2806-parscit.130908.xml (Conclusion)
*[...] The approach to the web as a genre repertoire in evolution and these preliminary findings **can** turn out to be useful when building web genre palettes or when designing new genre identification experiments. [...]*

(16) Source File: W06-2914-parscit.130908.xml (Conclusion)

[...] The results also suggest that word distributions themselves ***might*** *be a good candidate for capturing the thematic shifts of text and that SVM learning can play an important role in building an adaptable correlation. [...]*

The findings also support the initial hypothesis that I entertained about abstracts. Abstracts demonstrate the lowest frequency of modal auxiliaries and this can be explained by one of the abstract's intentions -- to persuade the reader of the significance of the conducted research. To convince them to read the full paper, scholars would generally avoid hedging markers that can evoke hesitation/uncertainty.

4.3.2.2 Modal Lexical Verbs

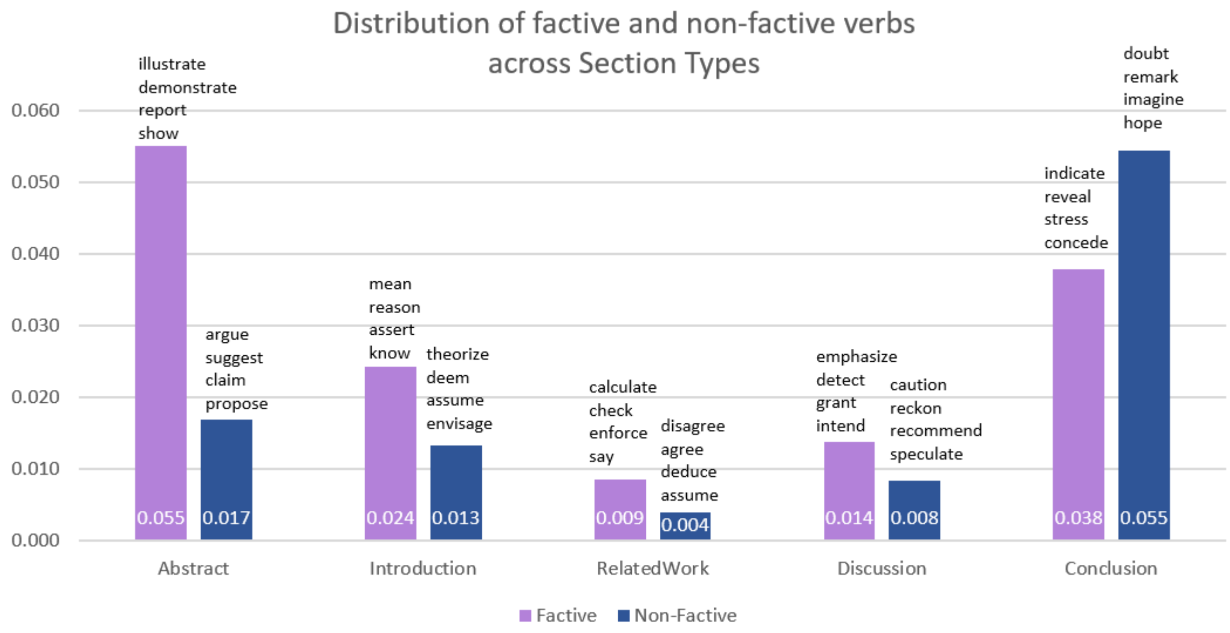


Figure 13 Distribution of factive and non-factive verbs across section types.

Figure 13 shows the distribution of factive and non-factive verbs across section types based on MI scores (the computed values are available on GitHub and on GUDe under **Linguistic Mechanisms > MI Scores > Modal Lexical Verbs**). The categorization of the verbs is based entirely on their use in the sections, i.e., the categorization is contextually bound. I understand factive and non-factive verbs to be verbs that combine with *that-complement clauses* and that grant factual or non-factual/evaluative status respectively to the statement made in the complement clause. It is also important to point out that the examples of factive and non-factive verbs co-occur with pronouns in the first person singular or plural, which can support Hyland's (1998) claim that self-mentions can also be considered hedging markers, especially when self-mentions occur in combination with non-factive verbs.

The graph shows the frequency of hedging marked by the presence of non-factive verbs, which can denote authors' personal evaluation of the research results and can also signal the degree of certainty with which they make their claims. The frequency of non-factive verbs was compared with the frequency of factive verbs in order to display the overall distribution of objective and evaluative language throughout the paper. The values shown in Figure 13 are based on the sum of the **M**utual **I**nformation (**MI**) scores of the factive and non-factive verbs from the first 50 section-representative examples.

The results show that abstracts demonstrate the highest frequency of factive verbs (0.055) and conclusions the highest frequency of non-factive ones (0.055). These findings confirm, on the one hand, the initial expectations and reflect, on the other, the author's intentions in the individual sections. Since abstracts are intended to persuade the reader to read the main body of the paper, authors would make sure that they avoid using verbs that can signal any form of hesitation/uncertainty. In fact, the high occurrence of factive verbs can be interpreted as a persuasion strategy on the part of the writers. By contrast, in conclusions, authors tend to use more non-factive verbs to indicate that the propositions they make in light of their research results are only one possible interpretation that they are proposing to the research community, and the piece of knowledge they present cannot be treated as a fact, but rather as a suggestion resulting from the adopted methods or the devised techniques. Another and perhaps stronger reason can be that the author speculates on what would follow from the results of the paper, how they might fit in with other research results, and which further experiments might be worth conducting.

The semantics of the four factive and non-factive verbs, which were most representative of a section, can also give us some insights into the communicative goal of the section types. For example, the four abstract-specific factive verbs include *illustrate, demonstrate, report,* and *show.* Authors use such verbs to indicate that the findings they report reveal the efficiency of the adopted approach/technique and provide examples of how these findings can be useful.

(17) Source File: P15-1019-parscit.130908.xml

Our model takes into account this information and precisely represents it using probabilistic topic distributions. We ***illustrate*** *that such information plays an important role in parameter estimation. [...]*

(18) Source File: W03-2602-parscit.130908.xml
We use simple noun chunking at the syntactic analysis stage and extract grammatical function information by pattern matching. Identifying subjects and objects is critical to salience calculations. We ***report*** *that this important subject-object distinction can be made reliably with our shallow approach.*

(19) Source file: W04-1216-parscit.130908.xml
While many systems have laboriously hand-coded rules for all kinds of word features, we ***show*** *that word similarity is a potential method to automatically get word formation, prefix, suffix and abbreviation information automatically from biomedical texts, as well as useful word distribution information.*

In addition to presenting briefly the facts around the analyzed phenomenon or the proposed approach/model, authors can also present, in a concise fashion, the results from the study and make claims on their basis or make suggestions about the possible effect these findings can have on the progress/development of a particular process/a model. This becomes evident from the use of the non-factive verbs *argue, suggest, claim,* and *propose*.

(20) Source File: W04-2117-parscit.130908.xml
We ***argue*** *that just as the mental lexicon exhibits various, possibly interwoven layers of networks, electronic LRs containing syntagmatic, morphological and phonological information need to be integrated into an associative electronic dictionary.*

(21) Source File: W05-1210-parscit.130908.xml
We ***suggest*** *that our models and annotation methods can serve as an evaluation scheme for entailment at these levels.*

In introductions, the most representative factive verbs are *mean, reason, assert* and *know*. The verb *mean* suggests that authors provide clarification of some concept or a novel method that they introduce to the research community. The verb *reason* indicates that they acknowledge what methods/techniques should be used in order for the desired effect to be achieved, or they recognize the conditions under which the tested model can perform better. They might also make *assertions* about the importance of the addressed questions and/or the significance of the conducted research.

(22) Source File: D09-1129-parscit.130908.xml

By multiple errors, we ***mean*** *that if we have n words in the input sentence, then we try to detect and correct at most n-1 errors.*

(23) Source File: W06-2506-parscit.130908.xml

We ***reason*** *that, if multiple meanings of an ambiguous word are activated when the stimulus is processed, then the elicited associates should reflect the ambiguity.*

(24) Source File: W14-3103-parscit.130908.xml

While visualizing human language is a broad subject, we apply Polya's dictum, and examine a pair of simpler questions for which we still lack an answer: • (1) what is in this corpus of documents? • (2) what is the relationship between these two corpora of documents? We ***assert*** *that addressing these two questions is a step towards creating visualizations of human language more suitable for exploratory data analysis.*

They may also *theorize* or make *assumptions* about the credibility of the findings/claims made by peer scholars. For example,

(25) Source file: W13-4012-parscit.130908.xml

In this paper, following Hearst (1997), we ***assume*** *that a text or a set of texts develop a main topic, exposing several subtopics as well. We also* ***assume*** *that a topic is a particular subject that we write about or discuss (Hovy, 2009), and subtopics are represented in pieces of text that cover different aspects of the main topic (Hearst, 1997; Hennig, 2009).*

By connecting their research topic to previous works, authors report on the information they *agree* or *disagree* with, they make *deductions* about the credibility of the content presented by peer scholars and comment on the *assumptions* made by them. This use of epistemic modal lexical verbs correlates with the accuracy-oriented function of hedges that Hyland introduces, because writers try to interpret the claims of other scholars as precisely as possible and stress the fact that these interpretations result from the content they have been exposed to.

(26) Source File: P13-1098-parscit.130908.xml
We ***disagree*** *with the arguments supporting the statement "you cannot predict elections with Twitter" (Gayo-Avello, 2012), as many times in the past actual voting intention polls have also failed to predict election outcomes, but we* ***agree*** *that most methods that have been proposed so far were not entirely generic.*

In addition, they contrast their approach with the ones previously described, by emphasizing the series of actions they have taken to carry out the study, which can lead to better performance/results -- *we* ***calculate*** *that..., we* ***check*** *that..., [...] and one way of comparing our results to theirs is to* ***say*** *that ....*

In discussions, authors *warn/caution* readers about the limitations/pitfalls of the adopted approach (e.g., data sparsity) or of the analyzed data that might have affected the accuracy of the presented scores or the performance of the tested model. By pointing out what has worked and what has not during their experiments, they make practical conclusions and give *recommendations* about the issues that should be considered or avoided in future studies. Considering the achieved results and presented facts, authors also make *speculations* about their possible interpretation or express their opinion on what can be the possible answers to the present questions. This personal evaluation of the results can be correlated with the reader-based function of hedges, with which, as Hyland suggests, authors implicitly invite readers to take part in the discussion and make their contribution.

(27) Source File: D14-1119-parscit.130908.xml
Although we are trying to mimic the situation in which we predict how an arbitrary user would vote on an arbitrary question, we ***caution*** *that the vote data we train and evaluate on was not obtained from a set of arbitrary SodaHead users. It consists only of votes from users who chose which questions they wanted to answer.*

(28) Source File: P07-1081-parscit.130908.xml
In order to alleviate some of these effects on the stability of word accuracy measures across corpora, we ***recommend*** *that at least four transliterators are used to construct a corpus.*

In this section, authors also *emphasize* in what ways their approach proved better or their model performed better than those previously described or they also acknowledge the improvements they get in their results or the improvements in the performance of the tested model under certain circumstances.

(29) Source File: P08-1076-parscit.130908.xml
*We **emphasize** that our model achieved these large improvements solely using unlabeled data as additional resources, without introducing a sophisticated model, deep feature engineering, handling external hand2[sic]Since CoNLL'00 shared task data has no development set, we[sic]divided the labeled training data into two distinct sets, 4/5 for training and the remainder for the development set, and determined the tunable parameters in preliminary experiments.*

(30) Source File: D10-1121-parscit.130908.xml
*We **see** that adding more data continues to increase the accuracy, and that accuracy is quite sensitive to the training data.*

In conclusions, authors formulate their claims on the basis of the content described in the discussion section. They *stress* the potential of the conducted research or explicitly express their *hopes* that their results and revelations would be a quality addition to the research field. Moreover, they might also talk about how they *imagine* future studies can benefit from their findings or express their doubts resulting from the limitations mentioned in the discussion and point out what aspects should be considered in future works.

(31) Source File: W98-1210-parscit.130908.xml
*Although the example presented in this paper used a natural language corpus, we **stress** that these techniques are suited to the analysis of all kinds of data.*

(32) Source File: W15-4612-parscit.130908.xml
*Our main aim in this paper was to show that experiments with discourse parsing can be done fairly easily using one of the many freely available sequential models. We **hope** that this method will make the task more accessible to researchers and help in moving towards a fully statistical and holistic approach to discourse parsing.*

(33) Source File: W04-2305-parscit.130908.xml
*Instead of selecting correct interpretations, we **imagine** that one could also use the proposed setup to decide which of a finite set of dialogue moves was performed by a speaker.*

The retrieved verb patterns (*I/we + verb + that-complement*) for each file per section type are available on GitHub and on GUDe under **Linguistic Mechanisms > supplements > vpattern.**

4.3.3 Evaluative and Non-evaluative Language in Section Types

In addition to the distribution of factive and non-factive verbs, an additional experiment was conducted to check the general distribution of evaluative and non-evaluative language across section types marked by the content parts-of-speech (POS). The frequencies are based again on the sum of the **M**utual **I**nformation (**MI**) scores of instances of nouns, verbs, adjectives, and adverbs (all computed raw frequencies and MI scores are available here, as well as on GUDe under **Linguistic Mechanisms > MI scores > POS**). These can denote either authors' personal assessment of the research results presented in the publication, signaled by sentiment-burdened words, as well as words marking the degree of certainty with which claims are made, or they can denote their objective presentation of the conducted research, indicated by words designating domain-specific characteristics, communicative activities, or discourse relations.

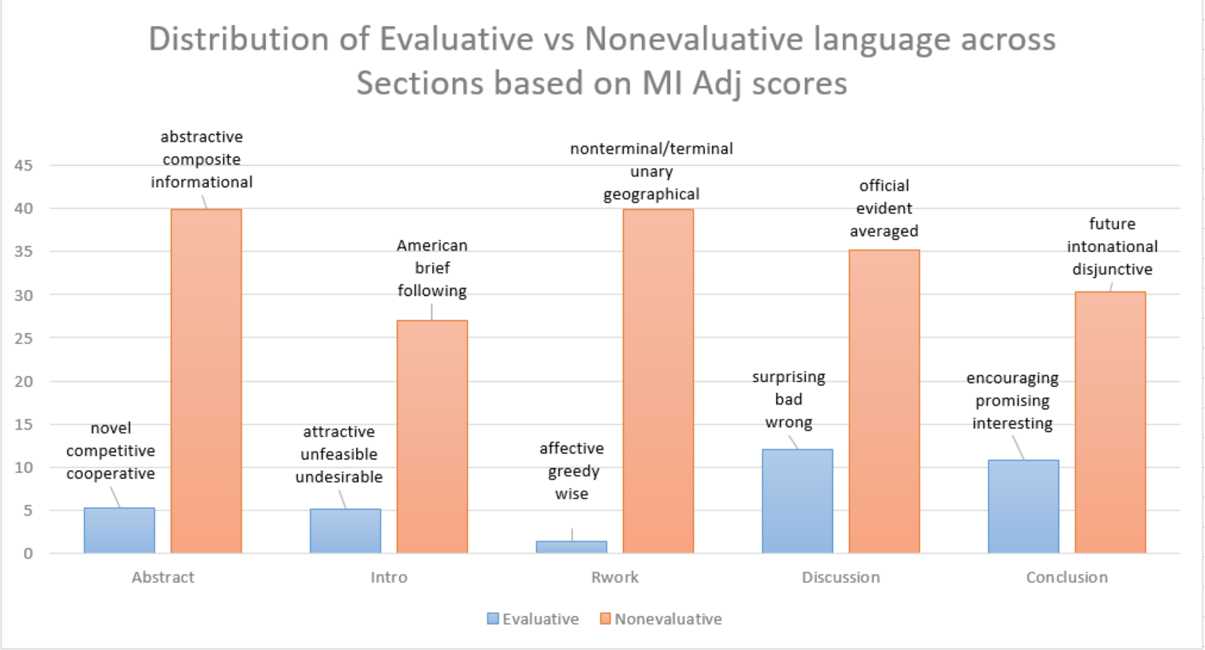


Figure 14 Distribution of Evaluative and Non-evaluative Language across Sections based on MI Adjective (Adj) Scores.

Figure 14, for example, shows the distribution of evaluative and non-evaluative adjectives across sections. The first 100 adjectives (ADJs), which are most representative of a section type, were classified either as evaluative or non-evaluative depending on whether they convey the authors' personal stance/ subjective interpretation of the research results or not. Examples of evaluative ADJs include *surprising, bad, novel, interesting, undesirable, encouraging*, etc., and examples of non-evaluative ones are *informational, brief, evident, disjunctive, etc.* Then the MI scores from the most representative instances of evaluative or non-evaluative adjectives were summed up to get the mean scores for each group per section.

By comparing the mean scores of the evaluative and non-evaluative examples per section, the results show that there is a considerable increase in the frequency of evaluative adjectives in discussions and conclusions with mean scores (11.99) and (10.82) respectively. The findings suggest that authors tend to evaluate the research results or present their stance on the success or failure of the presented research in discussions and conclusions. The increase in the frequency of evaluative language is accompanied by a decrease in the frequency of non-evaluative language in these sections. The frequency of evaluative ADJs is lowest in the related work section with a mean score (1.34), followed by the introduction section (5.07) and the abstract (5.25). This means that authors' expression seems to be more objective, devoid of any sentiments in abstracts, introductions, and related works. In abstracts, they present briefly the research milestones; in introductions, they give a detailed overview of the paper's content and in related works, they relate the current research to previous contributions in the discussed subject area. Abstracts and related works contain the highest frequencies of non-evaluative adjectives -- 39.92 and 39.90 respectively.

Furthermore, the 3 ADJs that appear on top of each group are the most typical/representative ones of a section type. The semantics of these most important ADJs plausibly correlate with the communicative goal of the section type. Since one goal of abstracts is to attract readers' attention to read the whole publication, the 3 evaluative adjectives with the highest MI scores (*novel, competitive* and *cooperative*) do not by accident express a positive sentiment. Their function is to extol the virtues of the research under discussion. In discussions and conclusions, the most important evaluative adjectives relate to the authors' evaluation of whether the current research was successful, unsuccessful, or contrary to their initial expectations. In the related work section, for example, the most important non-evaluative words are domain-specific examples (*nonterminal, unary, geographical*), which suggests that a good amount of jargon language is concentrated in this section. In the conclusion section, the adjective with the highest MI score is *future* since authors tend to talk about their future steps and how they intend to conduct their future research.

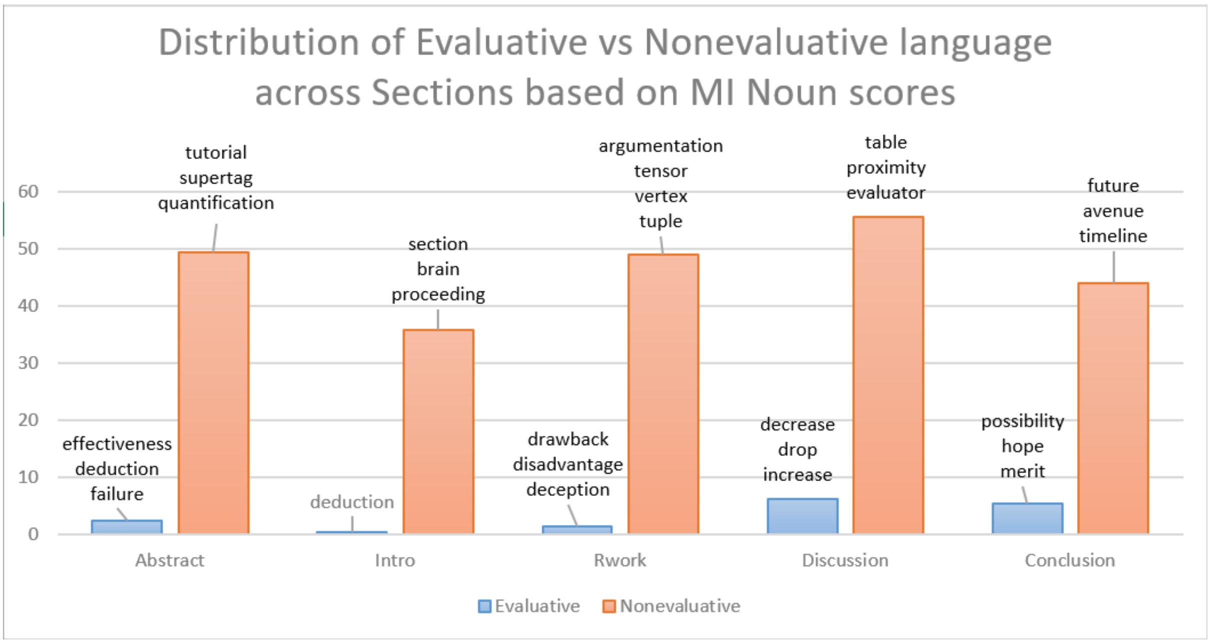


Figure 15 Distribution of Evaluative and Non-evaluative Language across Sections based on MI Noun Scores.

Similarly to Figure 14, Figure 15 also shows the increase in the frequency of evaluative language marked by nouns in discussions and conclusions with mean scores (6.24) and (5.41) respectively. The lowest frequencies of evaluative nouns are in introductions (0.42), followed by related works (1.32), and abstracts (2.33). Again, there seems to be a correlation between the most representative nouns of a group and the purpose of the section type. For example, the most important non-evaluative noun in the introduction section is *section* since authors give an overview of the paper's sections and the type of information that can be found in each of them. The most important non-evaluative noun in discussions is *table* since authors make references to various tables in which numerical results from empirical studies are stored. The words *future* and *avenue* are also correlated with the conclusion's purpose, namely, to present not only the authors' plans for the continuation of the research, but also the problems that should be addressed, how they should be approached, and the possible progress that can be made towards achieving a particular goal by adopting certain techniques/methods. When it comes to the evaluative nouns, the most important nouns in related works express a negative sentiment -- *drawback, disadvantage, deception*. A possible explanation for this is that in this section authors not only make references to previous studies, but they also emphasize the limitations or the disadvantages of these studies. This can serve as authors' motivation for the importance of the newly conducted research. In discussions, the nouns with the highest MI scores are *decrease,* *drop,* and *increase,* which describe the authors' analysis of the observed results. Finally, the nouns *possibility, hope,* and *merit* are the most representative evaluative nouns in conclusions, which can refer to the authors' hopes that the current study can contribute to the subject area and that the target audience can benefit from the presented study.

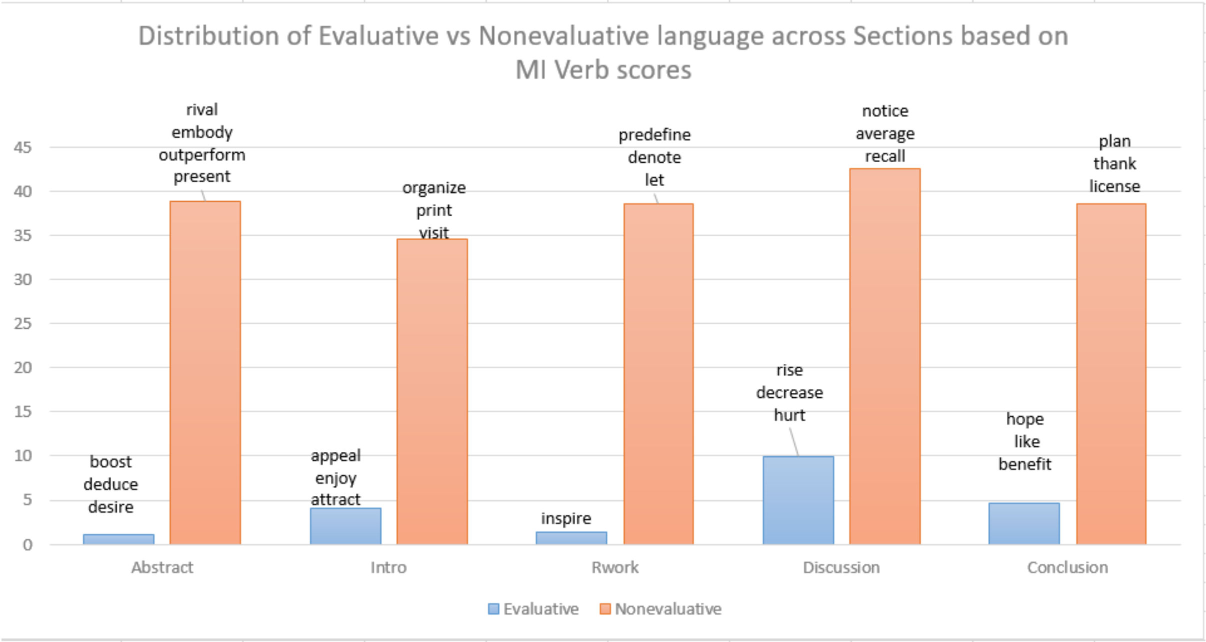


Figure 16 Distribution of Evaluative and Non-evaluative Language across Sections based on MI Verb Scores.

Similarly to evaluative adjectives and nouns in Figure 14 and Figure 15, Figure 16 shows that the frequency of evaluative verbs increases in the discussion section with a mean score of 9.90, followed by the conclusion section (4.69), and the introduction section (4.40). The lowest frequencies of evaluative language are again in the related work (1.35) and the abstract (1.05).

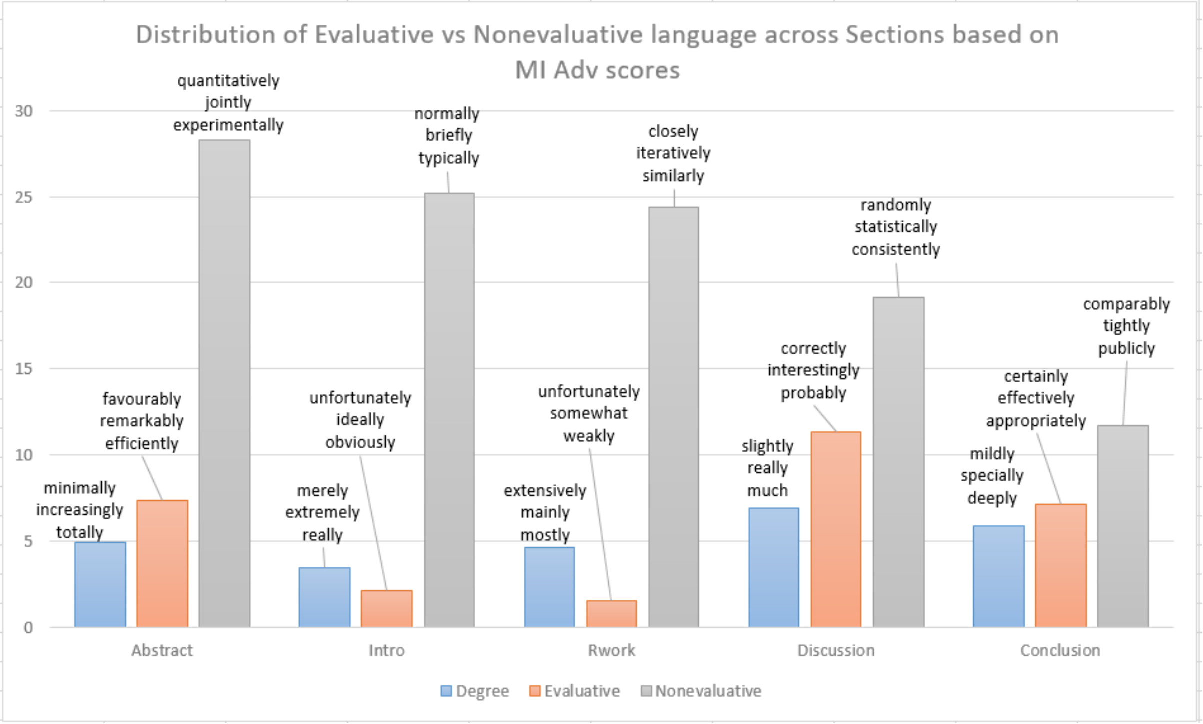


Figure 17 Distribution of Evaluative and Non-evaluative Language across Sections based on MI Adverb (Adv) Scores.

When it comes to adverbs, the frequency of evaluative adverbs and degree adverbs increases in discussions, and the frequency of non-evaluative adverbs decreases. This distribution can again be motivated by the communicative goals that authors intend to achieve in the different sections. Interestingly, abstracts demonstrate a comparatively high frequency of both evaluative and degree adverbs. The most representative adverbs for the abstract are *favourably*, *remarkably*, and *efficiently*, which are all with a positive sentiment. This can be explained by the fact that in abstracts, authors not only present the facts around their research, but they may also extol its virtues in order to engage readers to continue reading.

4.4 Cohesion

4.4.1 Coreference

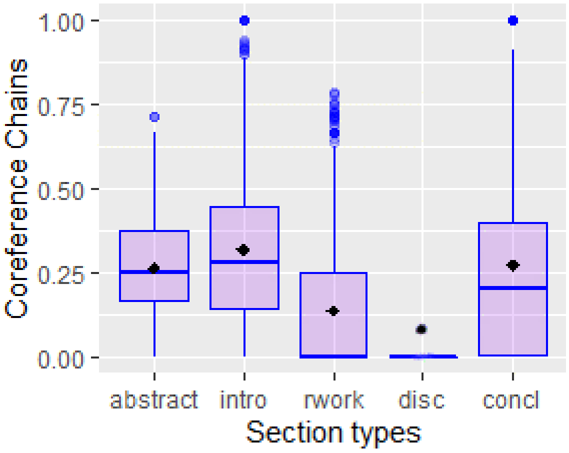


Figure 18 Distribution of Coreference Chains. Introductions demonstrate the highest frequency of coreference chains, discussions the lowest. In discussions, the median value is 0 and the mean is greater than 0 because more than half of the values in the sample are 0 and the rest are positive values.

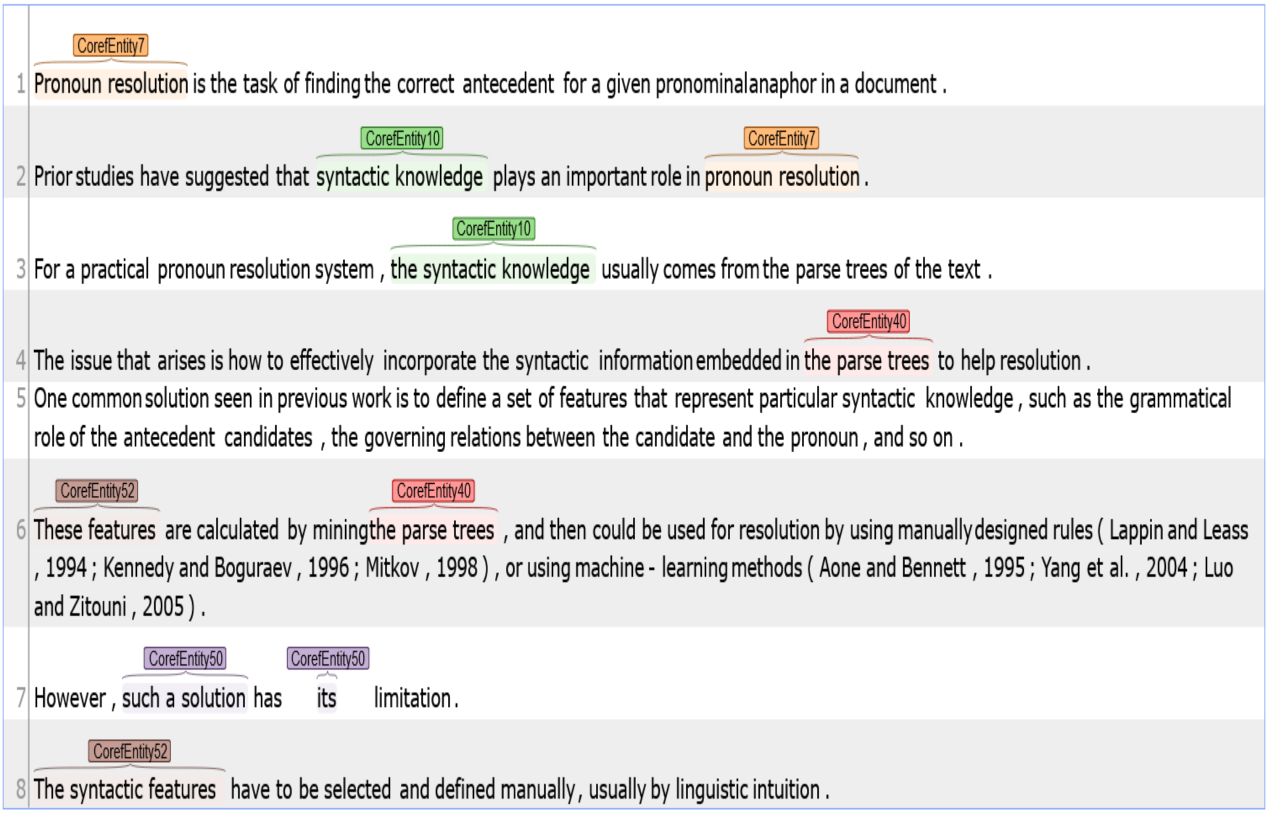
The graph shows the distribution of coreference chains across section types. Introductions demonstrate the highest frequency of coreference chains (0.43), followed by conclusions (0.34), abstracts (0.27), related works (0.24), and discussions (0.20). The result from the Wilcoxon signed-rank test shows that the overall difference in means is statistically significant with **p-value < 2.2e-16, V= 6550390**.

The results from the pairwise comparisons suggest that the differences in means between all pairs are statistically significant except for that between abstracts and conclusions.

Table 15 Coreference chains. Results from pairwise comparisons.

	abstract	introduction	related work	discussion
introduction	5.8e-14	-	-	-
related work	< 2e-16	< 2e-16	-	-
discussion	< 2e-16	< 2e-16	2.4e-16	
conclusion	1	1.4e-05	< 2e-16	< 2e-16

A possible explanation for the high frequency of coreference chains in introductions and related works can be the fact that in these sections authors make references to previous studies and their authors, present their research focus by juxtaposing their methods with previous ones, and as a result, they end up repeating topic-specific concepts in order to acquaint the reader with the research scope/problem. The anaphor in the anaphoric relations forming coreference chains can be realized by repeated proper nouns, personal pronouns, repeated NPs, or slightly modified NPs.



1 Pronoun resolution is the task of finding the correct antecedent for a given pronominalanaphor in a document .

2 Prior studies have suggested that syntactic knowledge plays an important role in pronoun resolution .

3 For a practical pronoun resolution system , the syntactic knowledge usually comes fromthe parse trees of the text .

4 The issue that arises is how to effectively incorporate the syntactic informationembedded in the parse trees to help resolution .

5 One commonsolution seen in previous work is to define a set of features that represent particular syntactic knowledge , such as the grammatical role of the antecedent candidates , the governing relations between the candidate and the pronoun , and so on .

6 These features are calculated by miningthe parse trees , and then could be used for resolution by using manuallydesigned rules (Lappin and Leass , 1994 ; Kennedy and Boguraev , 1996 ; Mitkov , 1998) , or using machine - learning methods (Aone and Bennett , 1995 ; Yang et al. , 2004 ; Luo and Zitouni , 2005) .

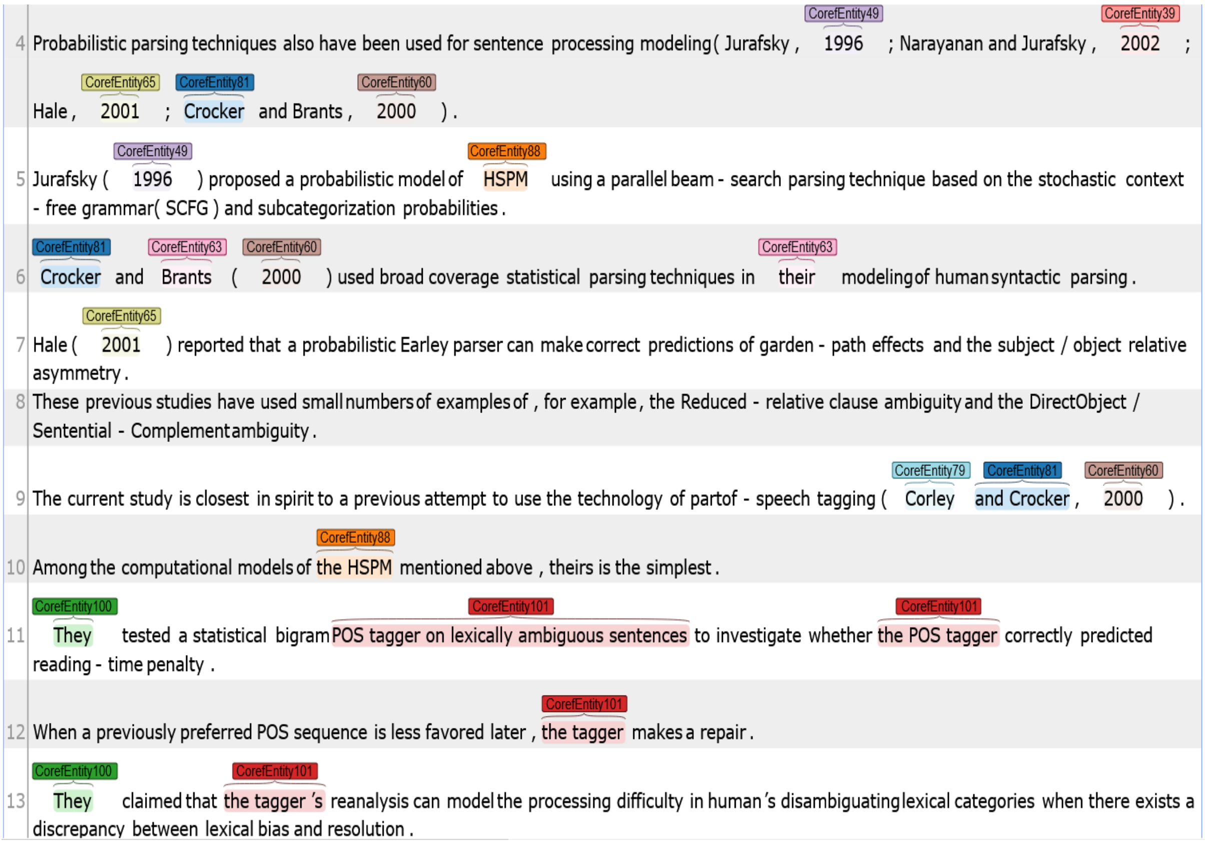
7 However , such a solution has its limitation .

8 The syntactic features have to be selected and defined manually , usually by linguistic intuition .

Figure 19 An excerpt of an introduction annotated with its coreference chains.

The example above (Figure 19) can give us insights into the type of information that is preserved throughout the text by looking closely at the annotated coreference chains in an introduction excerpt. The StanfordCoreNLP parser has detected 5 different coreference chains. In 3 of these chains, a coreferential relation is signaled by the repetition of subject-specific NPs: e.g., the repetition of *pronoun resolution*, which is mentioned twice and forms the first coreference chain; the repetition of *syntactic knowledge*, which is slightly modified by becoming definite later in the text *the syntactic knowledge*, which forms the second chain, and the repetition of *the parse trees*, which is mentioned again twice in the text and forms the third coreference chain. The fourth annotated coreference relation is established between the NPs *these features* and *the syntactic features*. The anaphor *the syntactic features* is modified by the adjective *syntactic*. Finally, the fifth relation is established between *such a solution* and the referring possessive pronoun *its*.

Considering the type of referents that appear in the example, it can be said that in introductions, authors tend to present in a detailed manner the nature of the research problem/adopted approach, in the context of CL, of the tested model. This becomes clear from the high frequency of mentions of topic-specific lexemes.



4 Probabilistic parsing techniques also have been used for sentence processing modeling(Jurafsky , 1996 ; Narayanan and Jurafsky , 2002 ; Hale , 2001 ; Crocker and Brants , 2000) .

5 Jurafsky (1996) proposed a probabilistic model of HSPM using a parallel beam - search parsing technique based on the stochastic context - free grammar(SCFG) and subcategorization probabilities .

6 Crocker and Brants (2000) used broad coverage statistical parsing techniques in their modeling of human syntactic parsing .

7 Hale (2001) reported that a probabilistic Earley parser can make correct predictions of garden - path effects and the subject / object relative asymmetry .

8 These previous studies have used small numbers of examples of , for example , the Reduced - relative clause ambiguity and the DirectObject / Sentential - Complement ambiguity .

9 The current study is closest in spirit to a previous attempt to use the technology of partof - speech tagging (Corley and Crocker , 2000) .

10 Among the computational models of the HSPM mentioned above , theirs is the simplest .

11 They tested a statistical bigram POS tagger on lexically ambiguous sentences to investigate whether the POS tagger correctly predicted reading - time penalty .

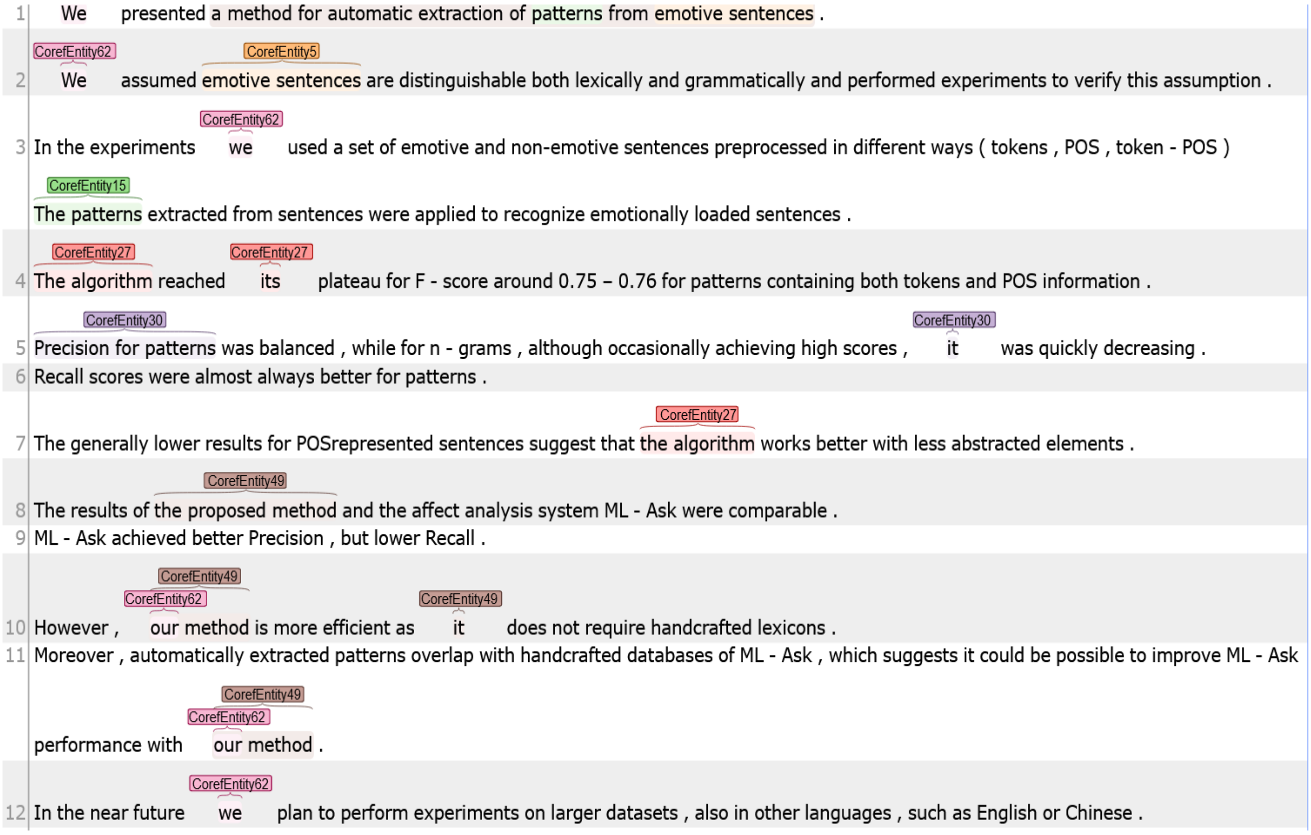
12 When a previously preferred POS sequence is less favored later , the tagger makes a repair .

13 They claimed that the tagger 's reanalysis can model the processing difficulty in human 's disambiguating lexical categories when there exists a discrepancy between lexical bias and resolution .

Figure 20 An excerpt of a related work section annotated with its coreference chains.

In the example excerpt from a related work section, the coreference chains are formed between entities referring to the authors of previously conducted studies in the research field or by the repetition of NPs referring to topic-specific concepts. The anaphors in the coreference relations established between entities referring to the authors of related works are realized by citations, by the use of the possessive pronoun *their* or the personal pronoun *they*. In the coreference chains formed between entities referring to subject-specific concepts, the anaphors take the form of repeated NPs or slightly modified NPs like in the following coreference chain: *POS tagger on lexically ambiguous sentences* (not identified correctly) - *the POS tagger* - *the tagger*.

The final example suggests that the identification of coreference chains can have certain limitations, which might have affected the accuracy of the coreference chains scores.



1 We presented a method for automatic extraction of patterns from emotive sentences .

2 We assumed emotive sentences are distinguishable both lexically and grammatically and performed experiments to verify this assumption .

3 In the experiments we used a set of emotive and non-emotive sentences preprocessed in different ways (tokens , POS , token - POS)
The patterns extracted from sentences were applied to recognize emotionally loaded sentences .

4 The algorithm reached its plateau for F - score around 0.75 – 0.76 for patterns containing both tokens and POS information .

5 Precision for patterns was balanced , while for n - grams , although occasionally achieving high scores , it was quickly decreasing .

6 Recall scores were almost always better for patterns .

7 The generally lower results for POSrepresented sentences suggest that the algorithm works better with less abstracted elements .

8 The results of the proposed method and the affect analysis system ML - Ask were comparable .

9 ML - Ask achieved better Precision , but lower Recall .

10 However , our method is more efficient as it does not require handcrafted lexicons .

11 Moreover , automatically extracted patterns overlap with handcrafted databases of ML - Ask , which suggests it could be possible to improve ML - Ask
performance with our method .

12 In the near future we plan to perform experiments on larger datasets , also in other languages , such as English or Chinese .

Figure 21 An excerpt of a conclusion annotated with its coreference chains.

Since, in conclusions, authors tend to make a summary of the conducted research, there can be a high frequency of backward-looking references to the approach the authors have adopted, the results they have achieved, the findings they have made, as well as a personal evaluation of their contribution to the field of research, and their plans for future actions. In the example above, the StanfordCore NLP Module has detected 5 different coreference chains. In the first one, the mentions are realized by repeated self-mentions *we*. Another coreference chain is formed between the NP *the algorithm* - *its* - *the algorithm*. In this case, the mentions of the algorithm under discussion are retained throughout the text by means of anaphors realized either by a proform (possessive pronoun) or a repeated NP, which refer back to the antecedent *the algorithm*. Another coreference chain is formed between the following entities *the proposed method* - *our method* - *it* - *our method*. Here the reference to the proposed method is retained throughout the text by means of slightly modified NPs *our method* or by a proform *it*. The final coreference chain is formed between the NP *Precision for patterns* and the proform *it*. The excerpt also demonstrates the frequent use of self-mentions in conclusions, which supports my findings so far.

4.4.2 Lexical Repetitions

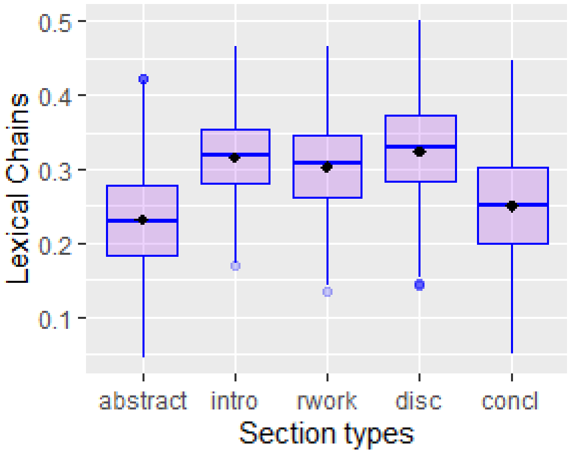


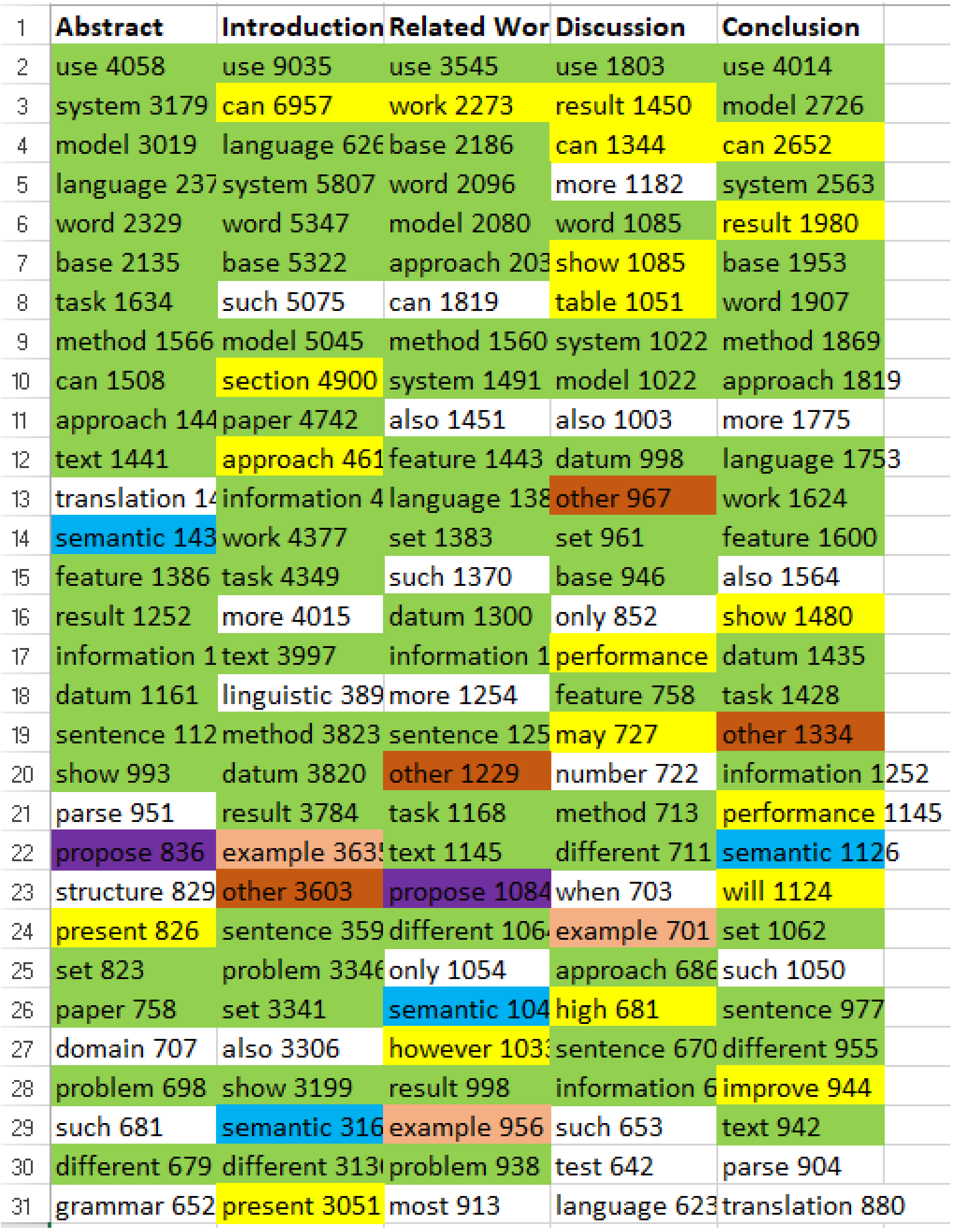
Figure 22 Distribution of Lexical Repetitions. Discussions demonstrate the highest frequency of lexical chains, abstracts the lowest.

Results show that the highest frequency of lexical chains is concentrated in discussions (0.32), introductions (0.31), and related works (0.30), followed by conclusions (0.25), and abstracts (0.23). The result from the Wilcoxon signed-rank test shows statistical significance with **p-value < 2.2e-16 and V=15587736**. The pairwise comparisons also confirm that the differences in means are statistically significant between all pairs.

Table 16 Lexical Repetitions. Results from the pairwise comparisons.

	abstract	introduction	related work	discussion
introduction	< 2e-16	-	-	-
related work	< 2e-16	8.9e-08	-	-
discussion	< 2e-16	0.0063	3.0e-14	-
conclusion	1.5e-09	< 2e-16	< 2e-16	< 2e-16

The quantitative assessment confirms the initial expectations and is in line with Wang & Zhang's (2019) findings: discussions and introductions demonstrate the highest frequency of lexical repetitions. This can also suggest that lexical repetition is one of the major devices that authors employ in these sections to signal cohesion/interconnectedness of the utterances. Yet, to examine what are the most frequently activated topics throughout the article and also to verify the hypothesis that the semantics of the lexical chains can be influenced by the section's communicative goal, the frequency of the most repeated words was computed.



1	Abstract	Introduction	Related Wor	Discussion	Conclusion
2	use 4058	use 9035	use 3545	use 1803	use 4014
3	system 3179	can 6957	work 2273	result 1450	model 2726
4	model 3019	language 626	base 2186	can 1344	can 2652
5	language 237	system 5807	word 2096	more 1182	system 2563
6	word 2329	word 5347	model 2080	word 1085	result 1980
7	base 2135	base 5322	approach 203	show 1085	base 1953
8	task 1634	such 5075	can 1819	table 1051	word 1907
9	method 1566	model 5045	method 1560	system 1022	method 1869
10	can 1508	section 4900	system 1491	model 1022	approach 1819
11	approach 144	paper 4742	also 1451	also 1003	more 1775
12	text 1441	approach 461	feature 1443	datum 998	language 1753
13	translation 14	information 4	language 138	other 967	work 1624
14	semantic 143	work 4377	set 1383	set 961	feature 1600
15	feature 1386	task 4349	such 1370	base 946	also 1564
16	result 1252	more 4015	datum 1300	only 852	show 1480
17	information 1	text 3997	information 1	performance	datum 1435
18	datum 1161	linguistic 389	more 1254	feature 758	task 1428
19	sentence 112	method 3823	sentence 125	may 727	other 1334
20	show 993	datum 3820	other 1229	number 722	information 1252
21	parse 951	result 3784	task 1168	method 713	performance 1145
22	propose 836	example 3635	text 1145	different 711	semantic 1126
23	structure 829	other 3603	propose 1084	when 703	will 1124
24	present 826	sentence 359	different 1064	example 701	set 1062
25	set 823	problem 3346	only 1054	approach 686	such 1050
26	paper 758	set 3341	semantic 104	high 681	sentence 977
27	domain 707	also 3306	however 1033	sentence 670	different 955
28	problem 698	show 3199	result 998	information 6	improve 944
29	such 681	semantic 316	example 956	such 653	text 942
30	different 679	different 313	problem 938	test 642	parse 904
31	grammar 652	present 3051	most 913	language 623	translation 880

Figure 23 The most highly ranked repetitions from the lexical chains for each section.

Figure 23 shows the first 30 most highly ranked repeated words from the lexical chains for each section. All word rankings are available on GitHub and on GUDe in the directory **Linguistic Mechanisms > statistical_analysis > Lexical Repetitions >**

lex_repetitions_ranked.xlsx. The words highlighted in green are the topics retained in all sections, e.g. *use, system, model, language, base, task, method, approach, datum,* etc.
These are all words that can be associated with the jargon of computational linguistics, and the fact that authors activate them in all sections can mean that they make sure that their argumentation does not lose focus at any stage of the writing process and that each topic brought up in a particular section can refer back to a topic mentioned in a preceding section or refer forward to a topic in a subsequent section. The identification of such recurrent topics throughout the article suggests that lexical repetitions play a significant role in the analysis of the topic persistence throughout the text.
The words highlighted in yellow are words whose occurrence in a particular section can be motivated by the section's purpose. To be more specific, the words *section* and *approach* can be viewed as introduction-specific lexemes. This can be explained by the fact that in introductions, authors typically give an overview of the article's content by explaining what type of information can be found in each section.

(34) Source file: J12-3005-parscit.130908.xml
*In the following, we first review related work (**Section** 2) and linguistic aspects of adjective classification (**Section 3**), then present the two acquisition experiments (**Sections 4 and 5**), and finish with a general discussion (**Section 6**) and some conclusions and directions for future research (**Section** 7).*

The same holds for the word *approach* since, in introductions, researchers either describe the approach they adopt for their experiment or test different approaches in order to find which of them leads to better results with respect to the research problem, as in the example below:

(35) Source file: J12-3005-parscit.130908.xml
*We first model polysemy in terms of independent classes to be separately acquired (e.g., an adjective with two senses ai and bi belongs to a class AB defined independently of classes A and B), and show that this model is not adequate. A second **approach**, which posits that polysemous adjectives simultaneously belong to more than one class (e.g., an adjective with two senses ai and bi belongs to both class A and class B), is more successful.*

Furthermore, the word *work* can be viewed as a lexeme associated with the related works since authors typically make references to previous works in order to explain the scope of their research.

(36) Source file: A00-1005-parscit.130908.xml
*As mentioned earlier, some customer service centers now allow users to say either the option number or a keyword from a list of options/descriptions. However, the only known **work** which automates part of a customer service center using natural language dialogue is the one by Chu-Carroll and Carpenter (1999).*

The words *result*, *show*, *table*, *performance*, etc. can be associated with discussions, alternatively with conclusions, in which authors typically present and comment on the results of the conducted study by referring to numeric scores organized in tables. In addition, in discussions/conclusions, authors tend to evaluate how well the tested model performed in comparison to previously tested models, so the word *performance* can also be considered a discussion/conclusion-specific word.

(37) Source file: A00-1007-parscit.130908.xml
*But we believe that the basic approach can be used also for multi-modal systems and other kinds of natural language dialogue systems. It is important to be aware of the limitations of the method, and how 'realistic' the produced **result** will be, compared to a dialogue with the final system.* (Discussion)

(38) Source file: A00-1036-parscit.130908.xml
***Table 1 shows** that for this task, the relaxation ranking passage retrieval algorithm without its supplementary knowledge sources (Recall II w/o knowledge) is roughly comparable in **performance** (42.9% versus 44.0% success rate) to a state-of-the-art commercial search engine (SearchIt) at the pure document retrieval task (neglecting the added benefit of locating the specific passages).* (Discussion)

(39) Source file: A00-1006-parscit.130908.xml
*We evaluated the translation **results** to see whether the impressions of the **results** improved or not.* (Conclusion)

(40) Source file: A00-1010-parscit.130908.xml

*To determine the **performance** of the system, we ran an informal experiment in which 11 different subjects called into the system and attempted to use it to solve a travel problem.* (Conclusion)

The occurrence of modal auxiliary verbs such as *can, may, will* can also be influenced by the section purpose. The frequency of these modals correlates with the distribution of modal auxiliaries that authors adopt when they hedge, as we saw in the section on hedging marked by Epistemic Modal Auxiliaries.

In addition, evaluative language, which is manifested by words loaded with either positive or negative sentiment like the adjective *high* in discussions and the verb *improve* in conclusions, can be considered a distinct lexical characteristic of discussions and conclusions. In discussions/conclusions, researchers typically express their personal evaluation of how accurately/well the tested model performed or whether the adopted approach has led to an improvement.

(41) Source File: A00-1038-parscit.130908.xml

*Further tests have shown that we can reach comparably **high** levels of accuracy for company topics when Entity Indexing is applied to financial, patents and public records sources.* (Discussion)

(42) Source File: A00-1011-parscit.130908.xml

*In this paper, we reported on a fast, portable, large-scale event and relation extraction system REES. To the best of our knowledge, this is the first attempt to develop an IE system which can extract such a wide range of relations and events with **high** accuracy. It performs particularly well on relation extraction, and it achieves 70% or higher F-Measure for 26 types of events already. In addition, the design of REES is highly portable for future addition of new relations and events.* (Conclusion)

(43) Source File: A00-2018-parscit.130908.xml

*[...] As shown in Figure 2, conditioning on this information gives a 0.6% **improvement**. We believe that this is mostly due to **improvements** in guessing the sub-constituent's pre-terminal and head.* (Discussion)

(44) Source File: A00-1028-parscit.130908.xml

[...] The current article demonstrates that a relatively simple pruning technique, employing the kind of reference corpus that is typically used for grammar development and thus often already available, can significantly ***improve*** *parsing performance. [...]* (Conclusion)

The distribution of evaluative language, across the section types, marked by content words, is analyzed in greater detail in the section on Evaluative and Non-evaluative language. The graphs also provide an elaboration of how the semantics of the lexemes can be mapped to the intentions of the section in which the words appear.

Considering the semantics of the most frequent lexical repetitions, it can be concluded that the communicative goal of each section interacts with the choice of vocabulary and their frequency of occurrence.

4.4.3 Explicit Connectives

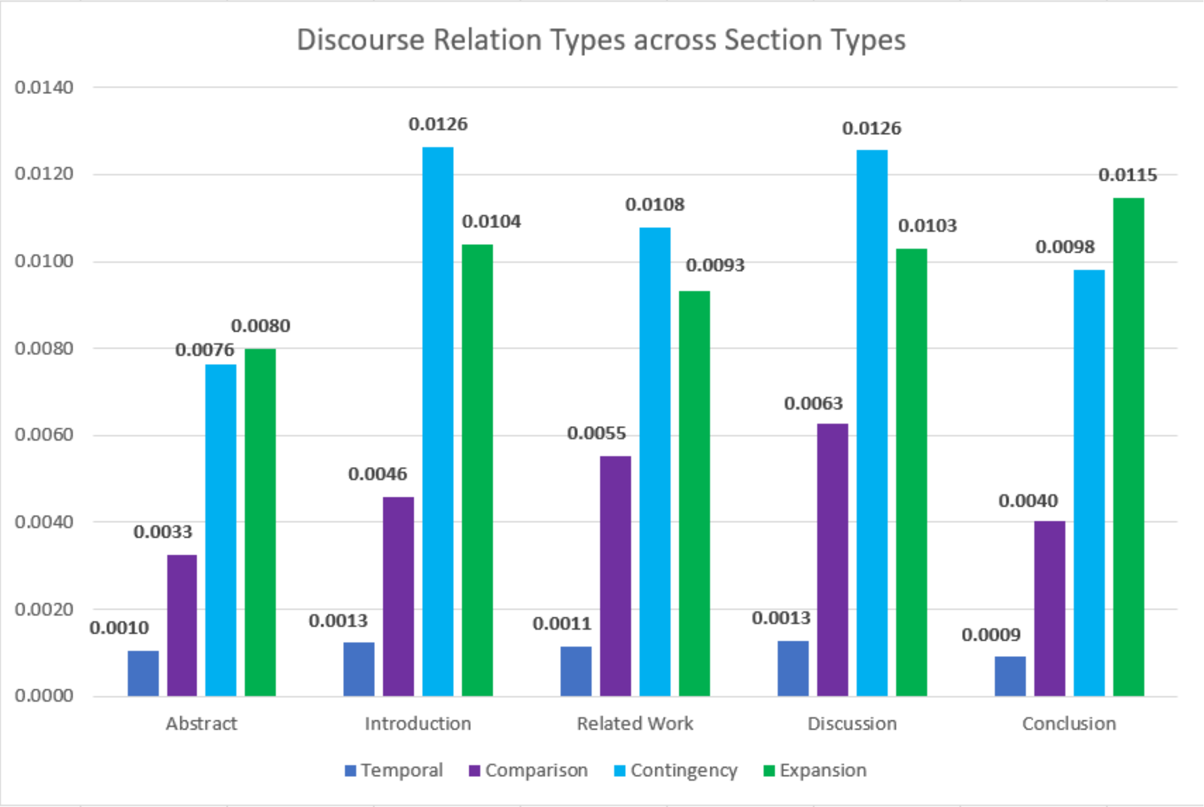


Figure 24 Distribution of the discourse relation types across the section types.

Figure 24 provides a bird's eye view of the distribution of all 4 relation types across the 5 section types. As expected, the frequency scores of the relation types vary inside each section and among the sections. The graph shows that abstracts, in general, demonstrate the lowest frequency of explicit connectives, which supports previous observations and the hypothesis that due to the section size, the existing relations are mostly inferred, rather than overtly expressed. The graph also shows that temporal relations demonstrate the lowest frequency across the sections, whereas contingency and expansion relations have the highest frequencies. To be more specific, in abstracts and conclusions, the expansion relation seems to be the predominant one. This can be explained by the fact that in abstracts, authors tend to present in a concise form the list of actions they have performed to achieve certain results. For this reason, in abstracts, expansion connectives such as *first, second/firstly, secondly, finally* can be expected like in the following example from an abstract:

(45) Source File: J79-1056-parscit.130908.xml
*This paper has three purposes: **firstly**, to describe how cage information is distributed in the preference semantics system of language understanding, and to show what practical use is made of that information. **Secondly**, to argue that that way of doing things has advantages over two alternatives: (a) putting all case information in one place, and (b) not using any case information at all, but only the names of English prepositions. **Thirdly**, I wish to... use the positions established earlier to counter some recent arguments by Charniak and others that the notion of case is not in fact functioning in any natural language understanding systems [...].*

What is more, the contingency relations are also prominent in abstracts, which can again be explained by their purpose. Abstracts can be viewed as a succinct summary of the research problem and a brief explanation of how to solve it. Authors would, therefore, use linkers that signal the cause-result connection. For example,

(46) Source File: P13-2070-parscit.130908.xml
*[...] Machine Transliteration is an essential task for many NLP applications. However, names and loan words typically originate from various languages, obey different transliteration rules, and **therefore** may benefit from being modeled independently. [..]*

In conclusions, authors reactivate the knowledge presented in the previous sections in order to evaluate the extent to which the initial expectations were met or in order to comment on the performance of the tested model by specifying its merits and shortcomings. Hence, one would expect a high frequency of connectives marking result relations such as *therefore, indeed, consequently* like in the example below:

(47) Source File: A00-2008-parscit.130908.xml
[..] While WordNet describes semantic relations between words, it does not recognize the conceptual schemas, i.e. frames, that mediate in these relations, and ***therefore*** *does not have the means to link arguments of predicating words with the semantic roles they express. COMLEX and NOMLEX provide detailed information about the syntactic frames in which verbs and nouns occur, but also lack a means to link syntactic arguments with semantic roles. FrameNet* ***therefore*** *provides information that complements major existing lexical resources.*

What is more, in conclusions, authors elaborate on their future steps and on how the current research can be extended and improved. This can explain the high frequency of expansion relations. Alternatively, conclusions may be teemed with examples that illustrate the claims authors make based on their observations, or authors may make a list of their findings:

(48) Source File: E14-4020-parscit.130908.xml
We have presented a simple method for generating surface-based patterns from parse trees which, besides avoiding the need for parsing test data, also increases extraction quality. By comparing supervised and unsupervised parsing, we ***furthermore*** *found that unsupervised parsing not only eliminates the dependency on expensive domain-specific training data, but also produces surface-based extraction patterns of increased quality. [...]*

Contingency relations are also predominant in introductions, related works, and discussions since authors make sure that they develop their ideas in a logical and well-motivated manner. Thus, in introductions, also in related works, they describe the research problem/the research aim and explain their motivation for taking up the challenge to solve the existing problem by using a particular model or a method. They may also explain why they believe their method can improve the performance of a tested model or that their model can outperform previously introduced models.

(49) Source File: A00-1033-parscit.130908.xml
Current information extraction (IE) systems are quite successful in efficient processing of large free text collections due to the fact that they can provide a partial understanding of specific types of text with a certain degree of partial accuracy using fast and robust language processing strategies (basically finite state technology). They have been "made sensitive" to certain key pieces of information and ***thereby*** *provide an easy means to skip text without deep analysis.* (Introduction)

(50) Source File: W98-0703-parscit.130908.xml
Some nodes don't have associated causes, ***so*** *they are just defined via unconditional probabilities (e.g., P(Cause2)). Taken together, the set of all the conditional and unconditional probabilities determine a joint distribution for all the nodes being modeled (e.g., P (Symptoml, SymptomN, C ousel[sic],...0 ause M)). Such global distributions are usually difficult to assess directly;* ***hence****, the Bayesian network provides a convenient formalism for specifying the same distribution via local distributions, under conditional independence assumptions.* (Related work)

In discussions, the cause-result relation is prevalent since they report on the actions that were performed and the consequences from their actions.

(51) Source File: A00-1038-parscit.130908.xml
[...]However, ***because*** *most of the place names we targeted lacked useful 280 internal structure, manual intervention was a part of creating all 800 definitions for places. [...]* (Discussion)

In addition, one can also witness an increase in the comparison type from the introduction section onwards. Comparison relations are most frequent in related works and discussions since, in related works, authors tend to juxtapose the current research with previous studies and to explain in what ways the current research approach differs from those previously proposed. In discussions, authors juxtapose expectations and results. They comment on the extent to which their initial expectations were similar to and/or different from the current results. Therefore, comparison expressions such as *Contrary to our expectations, By contrast/In contrast to*, *However* might be very prominent in these sections.

(52) Source File: W98-0701-parscit.130908.xml

*[...] These methods, **however**, focus on only two senses of a very limited number of nouns and therefore are not comparable with our approach.* (Related work)

(53) Source File: D12-1050-parscit.130908.xml

*[...] Again these methods differ in terms of how they implement compositionality: addition and multiplication are commutative and associative operations and thus ignore word order and, more generally, syntactic structure. **In contrast**, the recursive autoencoder is syntax-aware as it operates over a parse tree. **However**, the composed representations must be learned with a neural network. [...]*

The results from the Wilcoxon signed-rank test show that the differences in the means scores of the temporal, comparison, contingency, and expansion relations across the five section types are statistically significant.

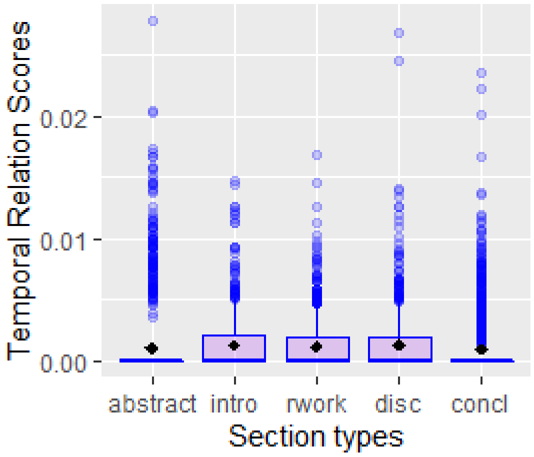


Figure 25 Distribution of Temporal Relations. Discussions demonstrate the highest frequency of temporal relations, abstracts and conclusions - the lowest. In all samples, the median is 0 and the mean is greater than 0. This is because more than half of the values in each sample are 0 and the rest are positive values.

The results show that there are occurrences of connectives marking temporal relations mostly in discussions (0.00128), introductions (0.00125), and related works (0.00114), whereas abstracts (0.00104) and conclusions (0.00090) demonstrate very low frequencies. The result from the statistical test shows significance with **p-value < 2.2e-16, V = 1549680**. The pairwise comparisons confirm that the difference in means is statistically significant between all pairs, except for that between introduction and related work, introduction and discussion, related work and discussion, as well as between abstract and conclusion.

Table 17 Temporal Relations. Results from the pairwise comparisons.

	abstract	introduction	related work	discussion
introduction	1.6e-07	-	-	-
related work	0.00031	1.00000	-	-
discussion	2.4e-08	1.00000	1.00000	-
conclusion	1.00000	6.2e-08	5.0e-05	3.2e-08

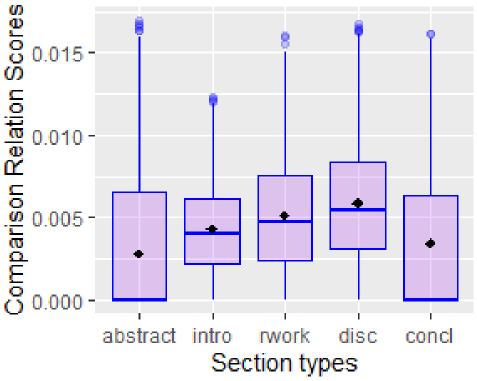


Figure 26 Distribution of Comparison Relations. Discussions demonstrate the highest frequency of comparison relations, abstracts and conclusions the lowest. Both in abstracts and in conclusions the median is 0 and the mean is greater than 0. This is because more than half of the values in both samples are 0 and the rest are positive values.

Similarly to temporal connectives, comparison connectives are also mostly used in discussions (0.0063), related works (0.0055), and introductions (0.0046). In abstracts (0.0033) and conclusions (0.0040), there is again a low frequency of this type of connectives. The results from the statistical test confirm that the differences in means are significant with **p-value < 2.2e-16**, **V = 7525260**. When compared pairwise, all differences in means between all pairs turned out to be statistically significant.

Table 18 Comparison Relations. Results from the pairwise comparisons.

	abstract	introduction	related work	discussion
introduction	< 2e-16	-	-	-
related work	< 2e-16	9.8e-08	-	-
discussion	< 2e-16	< 2e-16	8.0e-05	-
conclusion	0.0024	3.0e-07	< 2e-16	< 2e-16

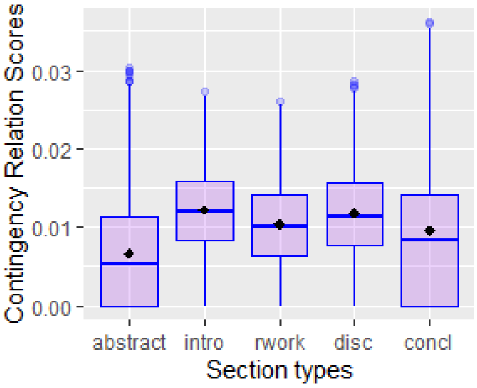


Figure 27 Distribution of Contingency Relations. Introductions and discussions demonstrate the highest frequency of contingency relations, abstracts the lowest.

In comparison to temporal and comparison connectives, there seems to be a higher occurrence of contingency connectives across all section types. They appear to be predominant in introductions (0.013), discussions (0.013), and related works (0.011), followed by conclusions (0.010) and abstracts (0.008). The result from the statistical test shows that again the differences in means are significant with **p-value < 2.2e-16, V = 10799628**. The pairwise comparisons confirm that the difference in means is significant between all pairs, except for that between introductions and discussions.

Table 19 Contingency Relations. Results from the pairwise comparisons.

	abstract	introduction	related work	discussion
introduction	< 2e-16	-	-	-
related work	< 2e-16	4.2e-14	-	-
discussion	< 2e-16	1	1.4e-07	-
conclusion	6.9e-10	< 2e-16	4.8e-05	< 2e-16

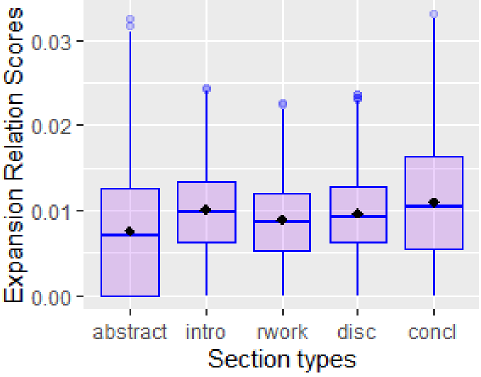


Figure 28 Distribution of Expansion Relations. Conclusions demonstrate the highest frequency of expansion relations, abstracts the lowest.

Similarly to contingency connectives, expansion connectives also demonstrate a high occurrence across all sections, more so in conclusions (0.012), discussions (0.010), and introductions (0.010). This time related works (0.009) and abstracts (0.008) seem to contain a lower occurrence of expansion connectives. The differences in means are statistically significant with **p-value < 2.2e-16**, **V = 11297881**. The pairwise comparisons show that

the only difference in means that is not statistically significant is between introductions and discussions.

Table 20 Expansion Relations. Results from the pairwise comparisons.

	abstract	introduction	related work	discussion
introduction	< 2e-16	-	-	-
related work	4.4e-07	6.5e-07	-	-
discussion	7.6e-16	1.00000	0.00026	-
conclusion	< 2e-16	0.06118	4.4e-10	0.00413

The following conclusions can be drawn from the quantitative assessment of the distribution of the cohesive devices across the section types.

First, the article sections differ in terms of the frequency of occurrence of coreference chains, lexical repetitions, and types of connectives.

Second, certain cohesive devices can be associated with a particular section type.

Table 21 Predominant cohesive devices in the section types.

Coreference	**Lexical Repetitions**	**Temporal Connectives**	**Comparison Connectives**	**Contingency Connectives**	**Expansion Connectives**
introductions	discussions	discussions	discussions	introductions discussions	conclusions
conclusions	introductions	introductions	related works		discussions introductions
abstracts	related works	related works	introductions	related works	
related works	conclusions	abstracts	conclusions	conclusions	related works
discussions	abstracts	conclusions	abstracts	abstracts	abstracts

Table 21 provides an overview of the analyzed cohesive devices and the sections in which they are predominant. The sections appear in a descending order, i.e., those on top demonstrate the highest frequency of a particular cohesive marker and those at the bottom the lowest frequency. The sections appearing in the same cell have the same frequency scores.

Lexical repetitions and **connectives** (temporal, comparison, contingency) seem to be the predominant cohesive mechanisms in **discussions** with which authors establish relations between blocks of information. **Coreferences** and **contingency connectives** seem to be predominant in **introductions**. **Expansion connectives** are mostly employed in **conclusions**. **Comparison connectives** appear mostly in **discussions**, but also in **related works**. Abstracts are the section-types in which authors seem to use fewer overt cohesive markers to signal relations that hold between entities. These relations, as was already proposed, are most probably inferred rather than explicitly expressed. Nevertheless, abstracts are not devoid of cohesive markers, authors tend to use coreference chains, in which anaphora is realized by personal or possessive pronouns.

Chapter 5 Sections and their Distinctive Features

In this section, I propose a mapping between section types and the analyzed linguistic features on the basis of their distribution across the five section types. For each section, I elicit a set of distinctive features, which either demonstrate a high or a low frequency of occurrence. For the corpus I investigated, the following facts have been supported empirically.

In **abstracts**, the features, which demonstrate the highest frequency of occurrence include NP Length, NP Complexity marked by the presence of adjective, noun, PP or past participle dependents of the noun head, as well as the highest frequency of non-evaluative adjectives and non-evaluative adverbs. By contrast, the features, with the lowest frequency of occurrence include hedging marked by passive voice, epistemic modal verbs, and epistemic lexical verbs, evaluative verbs, and cohesive devices such as lexical repetitions and explicit connectives signaling temporal, comparison, contingency, or expansion discourse relations.

In **introductions**, the features with the highest frequency include NP count, coreference, and explicit connectives marking contingency relations. The features with the lowest frequency include NP Complexity marked by past participle (VBN) modifiers, non-evaluative adjectives, evaluative and non-evaluative nouns, non-evaluative verbs, and degree adverbs.

Related works are associated with the highest frequency of NPs, passive voice, and non-evaluative adjectives, and the lowest frequency of self-mentions, hedging marked by modal lexical verbs, as well as the lowest frequency of evaluative adjectives, and evaluative adverbs.

Discussions demonstrate the highest frequency of modal auxiliary verbs as a marker of hedging, lexical repetitions, as well as evaluative adjectives, evaluative and non-evaluative nouns, evaluative and non-evaluative verbs, evaluative and degree adverbs. The features with the lowest frequency include NP count, NP length, and NP complexity marked by adjective, noun, and past participle (VBN) dependents, and coreference.

In **conclusions**, the features with the highest frequency include self-mentions, NP complexity marked by PP and past participle (VBN) dependents, modal auxiliary verbs, non-factive (modal lexical) verbs, and explicit connectives marking expansion relations. The features with the lowest frequency include passive voice, explicit connectives marking temporal relations, and non-evaluative adverbs.

These section-bound features can find good application in state-of-the-art automated tools for writing evaluation, text generation, or disinformation detection. To be more specific, such predefined features can be integrated as linguistic parameters/cues into tools assessing academic writing quality (e.g., *Coh-Metrix*[8]*, VisaS*[9], etc.). These parameters can improve the tool's functionality by enabling a detailed analysis of the typical linguistic mechanisms in a particular section type of research article and by providing informative feedback on whether the author of the text has achieved the communicative goal based on the presence or the absence of the target mechanisms. Furthermore, these text-type specific features can be used as predictors of the underlying lexical, syntactic, and discourse characteristics of a particular text type during the development of text generation tools (e.g., the transformer-based language model for text generation API based on the GT2-based model by *OpenAI*[10]). What is more, with the surge in disinformation dissemination through various information channels, there has been a dire need for automated tools for detecting deceptive and potentially harmful content in the digital space. Being exposed to overwhelming amounts of information daily, users may fail to pay close attention to the discourse characteristics such as source, genre-specific features that have to do with vocabulary, argumentation strategies, register, etc., which may help them decide if the content should be trusted or not. As a result, users can be deceived or tricked into acting in a particular way, which may have serious personal and/or social repercussions. Recent studies (Tomkins, 2019; Cohen, 2020; Sarts, 2020) stress the ever-increasing threats of the massive spread of disinformation. They address the consequences to which the uncontrollable dissemination of manipulated content with malicious intent can lead. For example, they point out that disinformation can pose threats to national security, instigate campaigns leading to social divisions (a most recent example of which is the division of society into supporters and opponents of the measures in the fight against the COVID-19 pandemic), or it can even inflame/intensify armed political conflicts. Due to the pressing need for tools that reliably detect and filter misleading and harmful information, recent studies (e.g., Tomkins 2019; Mahyoob et al., 2020) in the field of discourse analysis suggest that linguistic characteristics can play a significant role in the combat against disinformation. Tomkins and Mahyoob et al., for instance,

[8] http://cohmetrix.com/

[9] https://blog.studiumdigitale.uni-frankfurt.de/visas/software/

[10] Better Language Models and Their Implications (openai.com)

have analyzed datasets of authentic and fake news articles, and have found that these differ in the type of linguistic features and their frequency of occurrence.

Therefore, I strongly believe in the potential of the identification of text-type-specific linguistic characteristics (text-feature mapping). Such characteristics may be a stepping stone toward the development of more robust language-based or hybrid models for disinformation detection.

Although the applications of such distinctive linguistic features can be beneficial for those who would like to improve their writing skills by developing a conscious understanding of the underlying characteristics that make one section distinct from another, the possible risks of their misuse cannot be ignored. For example, when integrated as linguistic predictors in natural language generation tools, they may be used for undesirable purposes, e.g., for the automatic generation of research papers and other academic texts. To mitigate the risk of such unwanted practices, it is important that there is greater transparency and regulation of the purposes for which the tools are making use of such linguistic text-specific mechanisms.

Chapter 6 Conclusions and Future Steps

In the current study, I verified the hypothesis that the communicative goal of the section type places constraints on the choice of the linguistic mechanisms, their rhetorical functions, and their frequency of occurrence across five section types that appear in the same corpus of a research article. The results confirm, first, Grosz & Sidner's (1986) claims that the linguistic and the intentional structure of discourse are in constant interaction and that the discourse purpose influences the author's selection of lexical, syntactic, and discourse mechanisms during the production process. The results are also in line with von Stutterheim & Klein's (1989) conception of text structure and show that the question each section is produced to answer constrains the choice of referents and how they are retained throughout the individual sections but also between the sections.

I have hypothesized that abstracts are designed to convince the recipients to read the full content of the scientific paper by presenting in a concise, yet informative form the purpose, the methods, the results, and the possible contribution of the research. The results show that the presented information takes the form of long and syntactically complex NPs. Abstracts may not have a high frequency of NP occurrences, which can be explained by their compact sizes, but they seem to possess the longest token-based NPs in which the noun head takes different dependents -- adjectives, nouns, PPs, or past participles. In other words, the abstract's purpose encourages authors to provide as much information about their study as possible by producing information-loaded NPs. Such nominalizations make texts more abstract and technical and are believed to signal a high proficiency in writing. Another feature that can be associated with abstracts is the high frequency of non-evaluative, possibly domain-specific/technical adjectives and adverbs, which also add to the abstract's technicality and which can be viewed as a persuasion strategy that authors employ to emphasize the credibility and the high potential of the analyzed topic. What is more, the low frequency of any form of hedging can be considered a persuasion strategy. Authors avoid using any words or expressions which either express a lack of commitment to the presented information or invoke any thoughts of uncertainty/doubts in the recipients.

When it comes to text readability and coherence, it seems that relations tend to be inferred rather than overtly expressed, which accounts for the low frequency of explicit cohesive devices. On the basis of this interaction between the abstract's intention and the frequency of linguistic mechanisms, I conclude that abstracts are highly sophisticated pieces of writing in which information is presented densely and the relations between threads of meaning may not always be easy to process.

Unlike abstracts, introductions present in an extended form the scope of the research by specifying its aim and what methods and techniques were employed in the research in order for these aims to be achieved. Since introductions do not face the same size constraints as abstracts, there is less of a need for the information to be concentrated into complex NP structures. The NPs instead are shorter and less syntactically complex but with a higher frequency. What is more, relations between entities are overtly signaled by coreference chains in which the mentions are realized by the repetition of subject-specific concepts or personal pronouns. This can be accounted for by the fact that in introductions, authors may make references to peer scholars and their contributions to the topic under discussion. In addition, there is a low frequency of evaluative vocabulary, i.e., one would not expect authors to evaluate the efficiency or the performance of previously adopted methods/techniques but rather to objectively present their own methodology and how they plan to approach the research problem.

To motivate their choice of topic and to emphasize the potential of the research methodology, authors tend to relate their research focus to previously conducted studies in the related work section. They do so, on the one hand, in order to update the reader on the contributions that have been made in the particular field so far, to provide an overview of what other researchers have done, and what claims they have made based on their research outcomes. On the other hand, they do so in order to explain in what ways their study differs from, is a continuation of, or even is an improvement of the methods/approaches adopted previously. This explains the high frequency of NPs which mostly take the form of proper nouns (i.e., the names of the peer scholars), or referential pronouns. The high frequency of hedges marked by passive voice indicates that authors tend to detach themselves from the presented content or to indicate that they report the results from other studies as objectively and reliably as possible. The low frequency of evaluative vocabulary can be another indicator of the author’s attempt to keep his/her expression as technical and objective as possible.

In contrast to related works, discussions tend to be more subjective or personal since they aim to present the author's interpretation of the research results. This subjectivity is marked by the high frequency of epistemic modals, as well as the high frequency of evaluative nouns, adjectives, verbs, and adverbs. The high frequency of non-evaluative vocabulary also suggests that although authors convey their personal evaluation, they also tend to stay objective by activating some previous information regarding the techniques they have used and the series of actions they have performed in order to achieve the current results. The low frequency of long and complex NPs can be plausibly explained by the fact that in discussions, the focus is on the actions rather than on the performers, which are the authors themselves.

Finally, similarly to discussions, conclusions also possess the subjectivity/personal note and authors tend to employ the backward-looking strategy again. The author's voice is evident from the high frequency of self-mentions, which may have various rhetorical functions, e.g., to focus readers on the research milestones by reactivating readers' knowledge from the previous sections that has to do with the aim, the methods, the results, etc. Authors may also make self-references to signal that the claims they are making about the results are based on their judgments/observations/interpretations and they implicitly invite the research community to contribute to the analysis and interpretation of the research findings. Authors also use self-mentions to emphasize their contribution to the research field and to express their beliefs and hopes that this contribution can improve the performance of a model or solve an existing problem. The low frequency of passives also suggests that unlike related works, in conclusions, authors tend to adopt a more personal approach. Subjectivity is indicated by the high frequency of epistemic modal auxiliaries and epistemic lexical (non-factive) verbs. In addition, in conclusions authors adopt also a forward-looking strategy in the sense that they tend to discuss considerations that other researchers should keep in mind in future studies, or suggestions for research extension and improvement. Since authors also sum up what they have done and what results they have achieved, the relations between the different blocks of information are mostly linked by connectives marking expansion relations.

The current study also demonstrated that although the five sections answer different questions and have different communicative goals, they are mutually dependent and the content in one section is a natural continuation of the previous one. By recognizing the question that each constituent section is intended to answer, the reader will be able to recognize the question that the whole research article answers.

The current research can be expanded over different academic genres such as academic essays, dissertations, and presentations in different disciplines in the humanities or the hard sciences with an attempt to draw clearer linguistically motivated text-type and discipline-specific boundaries, which will assist both tutors and students involved in the academic writing process.

Finally, the identified distinctive features can be used for training language-based applications for text classification and/or text evaluation.

Appendices

Appendix 1: The Role of NLP in Overcoming the Challenges of Textual Data

Due to its inherent complexity, natural language-based data such as text, speech, etc. tends to pose various challenges when it comes to processing, analysis, and extraction of meaningful patterns. This section provides an overview of the methods and frameworks I used for automated text analysis and presents some of the issues I was confronted with during the text processing and feature collection so that these can be avoided in the future.

Textual data is unstructured (qualitative) data that, unlike structured (numerical) data, does not follow predefined models or schemes of organization. This generally impedes its management when it comes to its processing, searchability, and analysis. In addition, most statistical and machine learning models take numerical, not textual data as input. As a result, many data scientists who wish to draw inferences from a large number of texts (e.g., tweets, blog posts, articles, scientific papers, etc.) as efficiently as possible, feel discouraged to venture into working with such data and exploring its potential. Fortunately, owing to natural language processing (NLP), the analysis of large volumes of textual data within a particular domain and the extraction of patterns relevant to this domain do not look intimidating any longer (Sarkar, 2019). In brief terms, NLP is the practice of developing applications that facilitate the processing and analysis of natural language-based data (Sarkar 2019, p. 62).

In the current study, I applied NLP techniques mainly for the extraction of the five constituent sections and for the collection of the target linguistic features. To extract the five section types from the input XML files, I used the java Scanner method, which scans the XML content line by line and looks for the section names: *Abstract, Introduction, Related Work/Background, Discussion, Conclusion/Conclusions/Future Work.* If the line starts/ends with either of these section names, the program prints out all the lines containing the section content and ignores those lines containing metadata. Since the section names were stored within the **sectionHeader** tag, once the scanner detects a line that starts with this tag, it stops processing the document and does not print the next line because a new section starts.

An extract from a research article stored in XML:

```
<sectionHeader confidence="0.993867" genericHeader="abstract">
Abstract
</sectionHeader>
<bodyText confidence="0.999734857142857">
The paper describes a natural language based expert
system route advisor for the public bus transport
in Trondheim, Norway. The system is available on
the Internet,and has been intstalled at the bus com-
pany&apos;s web server since the beginning of 1999. The
system is bilingual, relying on an internal language
independent logic representation.
</bodyText>
<sectionHeader confidence="0.9988"
genericHeader="introduction">
1 Introduction
</sectionHeader>
```

The implementation logic, however, threw an exception and did not output the lines following the detected section name. What solved the problem was the addition of a condition to the while-loop (highlighted) which not only checks if the line following the detected section name starts with a **sectionHeader** tag, but it also makes sure that there is a next line in the document. As long as the scanner detects lines in the document, it scans them and looks for the predefined section keywords. If the line ends with or contains the name of the target section (in this case matching the keywords *Discussion/Discussions),* and as long as the current line does not start with the **sectionHeader** tag, and there are next lines in the document, it scans the next lines and prints out their content.

```
while (scan.hasNextLine()) {
     String myLine = scan.nextLine();
        if (myLine.endsWith("Discussion")
||myLine.endsWith("Discussions")
|| myLine.contains("Discussion")) {
```

```
while(!(myLine.startsWith("<sectionHeader"))&&scan.hasNextLine
())
{
    myLine = scan.nextLine(); }
```

All XML files were scanned and the different sections with their corresponding content were appended to separate tab-separated text (.txt) files in which each line corresponds to the file name ID and the respective section-type content. Thus, all detected abstracts were stored in one .txt file, all detected introductions in another .txt file, etc. This type of data organization facilitated the access, searchability, and collection of the linguistic features.

Then the extracted section types were further preprocessed by removing any tags or special symbols with the help of Python regular expressions.

The raw data was used as input to the pipelines of two NLP frameworks, which were leveraged for the processing and retrieval of the target features. The first one is the Stanford CoreNLP, which is a Java-based software for research purposes annotating the input texts with linguistic metadata such as part-of-speech tags, constituency analysis, coreference relations, sentiment analysis, etc. The CoreNLP pipeline was called on the section types individually and it was used for the extraction of coreference chains and sentence count, in particular. The pipeline outputs the total number of coreference chains identified in each file section.

The rest of the features were collected and extracted with Python and one of its most popular state-of-the-art libraries for NLP -- spaCy. Similarly to the CoreNLP pipeline, the spaCy pipeline processes texts, tokenizes them, assigns parts of speech to the tokens, lemmatizes, checks if the token is a number, word, or a punctuation mark,
assigns syntactic dependency labels by describing the relations between the constituents in terms of their syntactic functions (e.g., subjects, direct/indirect objects, predicative complements, etc.), identifies the head word in each constituent and its dependent words, and also segments the texts into sentences. All these spaCy annotators are powered by statistical models or rule-based matching methods. Statistical models enable the library to make predictions about the linguistic properties of the tokens as they appear in context. In contrast,

the rule-based matching engine allows the search and retrieval of exactly defined token sequences and phrases such as the passive voice pattern below:

```
passive_voice_rule =
[{'DEP':'nsubjpass'},{'DEP':'aux','OP':'*'},{'DEP':'auxpass'},{'TAG':'VBN'}]
```

Passive Voice in English is made up of a form of the auxiliary verb **be** and a **past participle**. The passive rule above searches for the following token pattern -- a noun subject + an auxiliary verb + auxiliary verb + past participle.
The 'DEP' attribute stands for the dependency relations that exist between the tokens. The second token takes an additional attribute 'OP', which stands for optional, and its value a '*', which allows patterns in which the auxiliary 'be' is preceded by another auxiliary, like in the following examples: *sentences have been extracted* or *constraints can be localized*, etc. Such rule-based patterns ensure a fine-grained search of the desired token sequences and return results with greater accuracy.

I used the statistical models to extract the parts of speech and the noun phrases (NPs), to measure the length of the NPs, and to analyze their internal structure in terms of the types of dependents that noun heads license. The dataset with all computed values for the different features per section is available on GitHub, as well as on GUDe under **Linguistic Mechanisms > dataset**.

For the extraction of the passive voice constructions, of the nouns taking past participle (VBN) dependents, as well as of the patterns of verbs taking a 1st person sg/plural subject and a *that-complement* clause, I resorted to a combination of the statistical model and the rule-based matching approach for maximum precision.

Appendix 2: Mutual Information and its Computation

Mutual Information (MI) is defined as an association metric between words. It measures the degree of association/proportion of a term (word/unigram), in the current study, of a content word (noun, verb, adjective, adverb) with a particular section type. The terms (words) with high MI scores per section are the words that the section type is more frequently associated with. For example, the term ***organize*** has a higher MI score in *introductions* than in the rest of the section types (see the computed frequencies and MI scores for all content words per section type here). This means that this term is an important word (a keyword) for *introductions*, i.e., it occurs more frequently in *introductions* than in the rest of the section types. The words with high MI scores, therefore, can be considered keywords (indicators) of a particular section type. All the extracted words occur in all five section types and the MI score of a word determines if it can be treated as a section-specific word or not.

The scores have been computed by dividing the frequency of a term in a particular section by its frequencies in total (the sum of all its frequencies in the five sections).

Appendix 3: List of Explicit Connectives

Table 22 List of explicit connectives.

Type of Relation	Temporal	Comparison	Contingency	Expansion
Connectives	“after” “afterwards" “before” “earlier” “later” “meanwhile” "next" "previously" "simultaneously” "shortly” “thereafter" "till" "until" "ultimately"	"although" "but" "conversely" "despite" "however" "instead" "nevertheless" "nonetheless" "rather" "regardless” "though" "whereas" "yet"	"accordingly" "as" "because" "consequently" "hence" “if" "indeed" "so" "thereby" "therefore" "thus"	"also" "alternatively" "besides" "else" “especially” "except" "finally" "first" "firstly" "further" "furthermore" "likewise" "moreover" "neither" "nor" "or" "otherwise" "overall" "plus" "second" "secondly" "separately" "similarly" "specifically"

Appendix 4: Documentation of the Scripts

Prerequisites for running the python scripts:

To run the python scripts, first, you need to:

1) download and install python 3.8 or later version

2) install the python library for NLP SpaCy using the following command:
   pip install spacy

3) download one of the spacy statistical models for English. The following command downloads the small statistical model for English:
   python -m spacy download en_core_web_sm

The python scripts are also available on GitHub, as well as on GUDe under **Linguistic Mechanisms > linguistic_features > venv > Scripts**.

Script 1: Tokens (tokens.py)

"""The script prints the total number of tokens per article section."""

```
import sys

"""The script below prints the number of tokens of abstracts per article.
To change the section type, replace "abstracts.txt" with
"introductions.txt", or
"relatedwork.txt", or
"discussions.txt", or
"conclusions.txt". """

f = open("abstracts.txt", "r", encoding="utf-8")
for line in f:
    if len(line) > 0:
        items = line.split("    ")
        if len(items) > 1:
            file_name = items[0]
            file_text = items[1]
            print(file_name)
            num_tokens = len(file_text.split(' '))
            if (num_tokens) > 0:
                print("Number of tokens:", num_tokens)
            else:
                print(0)
```

Script 2: NPs and Adj/Noun/PP dependents (nps.py)

```
"""This script prints the NP occurrences and their total number per article section.
It also prints the NPs in which the noun head takes dependents (Adj/Noun/PP) and
counts the number of these NPs."""

import sys
# Importing the library for NLP
import spacy

# Loading the statistical model and storing it in the nlp object.
# The statistical model enables spacy to make predictions about
# the linguistic attributes of the tokens in context.
nlp = spacy.load("en_core_web_sm")

# Opening and reading the raw data with the extracted section types.
# Here it is set to abstracts. To run the script on the rest of the section types, simply replace
# the file name with:
# "introductions.txt" -- introductions
# "relatedwork.txt" -- related works
# "discussions.txt" -- discussions
# "conclusions.txt" -- conclusions.

f = open("abstracts.txt", "r", encoding="utf-8")

# Accessing the textual content of the tab-separated file and storing it in the "file_text"
variable
for line in f:
    if len(line) > 0:
        items = line.split("   ")
        if len(items) > 1:
            file_name = items[0]
            file_text = items[1]

            # Converting the text to lower case
            low_case = file_text.lower()

            # Calling the nlp object on the sections to be processed. The nlp object contains the
            # processing pipeline and the language-specific rules for tokenization.
            doc = nlp(low_case)

            # Initializing an empty list to which all NP occurrences per article section will be
appended.
            NPs = []
            # NP_lengths= []

            # An empty list to which all NPs containing adjective dependents will be appended.
            ADJ_DEP = []
            # An empty list to which all NPs containing noun dependents will be appended.
            NOUN_DEP = []
```

```
# An empty list to which all NPs containing PP dependents will be appended.
PREP_DEP = []

# Iterate over each constituent in the document and append to the empty list
# the NP (noun constituents). Then count the total number of NPs in each list.
for chunk in doc.noun_chunks:
    NPs.append(chunk)
# print("NPs: ", NPs)
for chunk in NPs:
    for token in chunk:
        for child in token.children:
            if child.pos_ == "ADJ":
                ADJ_DEP.append(chunk)
            elif child.pos_ == "NOUN":
                NOUN_DEP.append(chunk)
            elif child.pos_ == "ADP":
                PREP_DEP.append(chunk)
print("File name ", file_name)
print("NPs: ", NPs)
print("Total number of NPs:", len(NPs))
print("NPs containing ADJs:", ADJ_DEP)
print("Total # NPs containing ADJs: ", len(ADJ_DEP))
print("NPs containing NOUN_DEP:", NOUN_DEP)
print("Total # NPs containing NOUN_DEP: ", len(NOUN_DEP))
print("NPs containing PPs: ", PREP_DEP)
print("Total # NPs containing PPs: ", len(PREP_DEP))
```

Script 3: NPs and VBN dependents (noun_VBN_dependent.py)

```
"""The script retrieves the occurrences of nouns taking a past participle (VBN)
dependent using spacy's rule-based matching engine and then counts their number per
article section.The rule-based matching engine allows the search and the retrieval of
exactly defined token sequences and phrases such as the passive voice pattern below."""

import spacy

# Importing the rule-based matching engine
from spacy.matcher import Matcher

from collections import Counter
nlp = spacy.load("en_core_web_sm")
matcher = Matcher(nlp.vocab)

# Opening and reading the raw data with the extracted section types.
# Here it is set to abstracts. To run the script on the rest of the section types, simply replace
# the file name with:
# "introductions.txt" -- introductions
# "relatedwork.txt" -- related works
# "discussions.txt" -- discussions
# "conclusions.txt" -- conclusions.

f = open("abstracts.txt", "r", encoding="utf-8")
line_cnt = 0
for line in f:
    if len(line) > 0:
        items = line.split("   ")
        if len(items) > 1:
            file_name = items[0]
            file_text = items[1]
            lower_case = file_text.lower()
            doc = nlp(lower_case)
            sents = list(doc.sents)
            all_matches = []
            np_rule = [
                # {'POS': 'DET', 'OP': '*'},
                # {'POS':'ADJ', 'OP': '*'},
                {'POS': 'NOUN'},

                {'TAG': 'VBN', 'OP': '+'}
            ]
            matcher.add('NP Rule', None, np_rule)
            matches = matcher(doc)
            print(file_name)
            for match_id, start, end in matches:
                matched_span = doc[start:end]
                # print(matched_span.text)
                all_matches.append(matched_span.text)
```

```
pastpart_mod_count = len(all_matches)
print("Nouns taking a VBN dependent: ", all_matches)
print("Total number of NPs containing a VBN: ", pastpart_mod_count)

   # Operators and quantifiers
   #  'OP': '!' Negation: match 0 items
   #  'OP': '?' Optional: match 0 or 1 times
   #  'OP': '+' Match 1 or more times
   #  'OP': '*' Match 0 or more times
```

Script 4: NP Length (np_length.py)

```
"""The script prints the token-based NP length and the sum of all NP lengths for each
article section."""

import sys
# Importing the library for NLP
import spacy

# Loading the statistical model and storing it in the nlp object.
# The statistical model enables spacy to make predictions about
# the linguistic attributes of the tokens in context.
nlp = spacy.load("en_core_web_sm")

# Opening and reading the raw data with the extracted section types.
# Here it is set to abstracts. To run the script on the rest of the section types, simply replace
# the file name with:
# "introductions.txt" -- introductions
# "relatedwork.txt" -- related works
# "discussions.txt" -- discussions
# "conclusions.txt" -- conclusions.
f = open("abstracts.txt", "r", encoding="utf-8")

# Accessing the textual content of the tab-separated file and storing it in the "file_text"
variable
for line in f:
    if len(line) > 0:
        items = line.split("   ")
        if len(items) > 1:
            file_name = items[0]
            file_text = items[1]

            # Converting the text to lower case
            low_case = file_text.lower()

            # Calling the nlp object on the texts to be processed. The nlp object contains the
            # processing pipeline and the language-specific rules for tokenization.
            doc = nlp(low_case)

            # Initializing an empty list to which all NP occurrences per  will be appended.
            NPs = []

            # Initializing an empty list to which the lengths of the individual NP chunks
            # per article section will be appended.
            NP_lengths= []

            # Iterate over each constituent in the document and append to the empty list
            # the NP (noun constituents). Then count the total number of NPs in each list.
            for chunk in doc.noun_chunks:
                NPs.append(chunk)
```

```
for chunk in NPs:
    for token in chunk:
        chunk_length = len(chunk)
    NP_lengths.append(chunk_length)

print("File name ", file_name)
# print("NPs: ", NPs)
print("NP lengths: ", NP_lengths)
print("Sum of the NP lengths: ", sum(NP_lengths))
```

Script 5: Passive Voice (passive.py)

```
"""The script retrieves occurrences of passive voice using spacy's rule-based matching
engine.The rule-based matching engine allows the search and the retrieval of exactly
defined token sequences and phrases such as the passive voice pattern below.
The pattern is adapted from
https://gist.github.com/armsp/30c2c1e19a0f1660944303cf079f831a."""

import spacy
from spacy.matcher import Matcher
from collections import Counter
nlp = spacy.load("en_core_web_sm")
matcher = Matcher(nlp.vocab)

# Opening and reading the raw data with the extracted section types.
# Here it is set to conclusions. To run the script on the rest of the section types, simply replace
# the file name with:
# "abstracts.txt" -- abstracts
# "introductions.txt" -- introductions
# "relatedwork.txt" -- related works
# "discussions.txt" -- discussions

f = open("conclusions.txt", "r", encoding="utf-8")
line_cnt = 0
for line in f:
    if len(line) > 0:
        items = line.split("    ")
        if len(items) > 1:
            file_name = items[0]
            file_text = items[1]
            doc = nlp(file_text)
            sents = list(doc.sents)
            all_matches = []
            # Passive Voice Pattern
            passive_rule =
[{'DEP':'nsubjpass'},{'DEP':'aux','OP':'*'},{'DEP':'auxpass'},{'TAG':'VBN'}]
            matcher.add('Passive',None,passive_rule)
            matches = matcher(doc)
            print(file_name)
            for match_id, start, end in matches:
                matched_span = doc[start:end]
                # print(matched_span.text)
                all_matches.append(matched_span.text)
            passives_count = len(all_matches)
            print("Passives found in the section: ", all_matches)
            print("Total number of passive constructions: ", passives_count)
```

Script 6: Frequency of past participles extracted from passive constructions (freq_VBN_in passives.py)

```
"""The script processes all extracted past participle (VBN) from
the passive voice constructions and retrieves the most frequent VBNs per section
type."""

import logging
import spacy
logging.basicConfig(level=logging.DEBUG)
from spacy.matcher import Matcher

# Setting the file containing the passive_voice matches per section type.
# Here it is set to the passive voice matches found in the conclusion sections.
# To get the most frequent VBN from the rest of the sections, simply replace the
FILE_NAME with :
# 'passive_matches_abstract.txt' -- abstracts,
# 'passive_matches_intros.txt' -- introductions,
# 'passive_matches_rworks.txt' -- related works,
# 'passive_matches_discs.txt' -- discussions.

FILE_NAME = 'passive_matches_concls.txt'

# Set this to the maximum number of files you want to process.
LIMIT = 8000

def count_frequency(my_list):
    # Creating an empty dictionary
    freq = {}
    for item in my_list:
        if item in freq:
            freq[item] += 1
        else:
            freq[item] = 1

    for k in sorted(freq, key=freq.get, reverse=True):
        print(k, freq[k])


def main():
    logging.info('Running freq_counter.py...')
    # Install with:
    # python -m spacy download en_core_web_sm
    nlp = spacy.load("en_core_web_sm")
    intros = open(FILE_NAME, 'r', encoding="utf-8")
    lines = intros.readlines()
    lemma_list = []
    all_matches = []
    file_cnt = 1
    for line in lines:
        split = line.strip().split('\t')
        if len(split) > 1:
            # Get an intro.
            an_intro = line[line.index("\t"):].strip()
            # print(an_intro)
            # Pos tag it.
            doc = nlp(an_intro)
            matcher = Matcher(nlp.vocab)
            sents = list(doc.sents)
            passive_rule = [{'TAG': 'VBN'}]
            matcher.add('Passive', None, passive_rule)
            matches = matcher(doc)
            for match_id, start, end in matches:
                matched_span = doc[start:end]
                # print(matched_span.text)
                all_matches.append(matched_span.text)
            # print("Processed file num:", file_cnt)
            file_cnt += 1
            if file_cnt == LIMIT:
                break
    count_frequency(all_matches)

if __name__ == "__main__":
    main()
```

Script 7: Modal Auxiliary Verbs (modal_aux_verbs.py)

"""The script prints the occurrences of modal auxiliary verbs and counts their number per article section."""

```
import sys
import spacy
from collections import Counter
nlp = spacy.load("en_core_web_lg")
f = open("abstracts.txt", "r", encoding="utf-8")
for line in f:
    if len(line) > 0:
        items = line.split("    ")
        if len(items) > 1:
            file_name = items[0]
            file_text = items[1]
            lower_text = file_text.lower()
            doc = nlp(lower_text)
            modals = ["might", "may", "could", "should", "can", "would", "will", "shall"]
            hedges = []
            for token in doc:
                if token.text == token.text in modals:
                    hedges.append(token.text)
            mod_aux_count = len(hedges)
            print(file_name)
            print("Modal auxiliaries found in the section: ", hedges)
            print("Total number of modal auxiliaries: ", mod_aux_count)
```

Script 8: Verbs co-occurring with pronominal subject and a *that*-complement (vpattern.py)

"""The script prints the occurrences of the pattern (I/we + verb + that-compl) per article section and then prints their total number."""

```
import spacy
from spacy.matcher import Matcher
from collections import Counter
nlp = spacy.load("en_core_web_sm")
matcher = Matcher(nlp.vocab)

# Opening and reading the raw data with the extracted section types.
# Here it is set to discussions. To run the script on the rest of the section types, simply replace
# the file name with:
# "abstracts.txt" -- abstracts
# "introductions.txt" -- introductions
# "relatedwork.txt" -- related works
# "conclusions.txt" -- conclusions.
f = open("discussions.txt", "r", encoding="utf-8")
line_cnt = 0
for line in f:
    if len(line) > 0:
        items = line.split("   ")
        if len(items) > 1:
            file_name = items[0]
            file_text = items[1]
            lower_case = file_text.lower()
            doc = nlp(lower_case)
            sents = list(doc.sents)
            all_matches = []
            verb_rule1 = [{'DEP': 'nsubj', 'POS': 'PRON', 'TEXT': 'we'},
                          {'POS': 'VERB'},
                          {'LEMMA': 'that', 'POS': 'SCONJ'}]
            verb_rule2 = [{'DEP': 'nsubj', 'POS': 'PRON', 'TEXT': 'i'},
                           {'POS': 'VERB'},
                           {'LEMMA': 'that', 'POS': 'SCONJ'}]
            matcher.add('Verb Rule1', None, verb_rule1)
            matcher.add('Verb Rule2', None, verb_rule2)
            matches = matcher(doc)
            for match_id, start, end in matches:
                matched_span = doc[start:end]
                # print(matched_span.text)
                all_matches.append(matched_span.text)
            vpattern_count = len(all_matches)
            print(file_name)
            print("Verb patterns: ", all_matches)
            print("Total number of verb patterns: ", vpattern_count)
                # Operators and quantifiers
                #  'OP': '!' Negation: match 0 items
                #  'OP': '?' Optional: match 0 or 1 times
                #  'OP': '+' Match 1 or more times
                #  'OP': '*' Match 0 or more times
```

Script 9: Frequency of verbs extracted from the verb pattern (I/we + verb + that-compl) (**freq_vpattern.py**)

```
"""The script processes the extracted verb patterns found in the article sections
(pronoun+verb+that-clause) and prints the most frequent verbs per section type."""


import logging
import spacy
logging.basicConfig(level=logging.DEBUG)
from spacy.matcher import Matcher

# Setting the file containing the extracted verb patterns (pronoun+verb+that-clause).
# Here it is set to the verb patterns found in the conclusion sections.
# To get the most frequent verbs from the rest of the sections, simply replace the
FILE_NAME with :
# 'vpatternText_abstracts.txt' -- abstracts,
# 'vpatternText_intros.txt' -- introductions,
# 'vpatternText_rworks.txt' -- related works,
# 'vpatternText_discs.txt' -- conclusions.
FILE_NAME = 'vpatternText_concls.txt'

# Set this to the maximum number of files you want to process.
LIMIT = 8000

# Defining a frequency counter function


def count_frequency(my_list):
    # Creating an empty dictionary
    freq = {}
    for item in my_list:
        if item in freq:
            freq[item] += 1
        else:
            freq[item] = 1

    for k in sorted(freq, key=freq.get, reverse=True):
        print(k, freq[k])


def main():
    logging.info('Running freq_counter.py...')
    # Install with:
    # python -m spacy download en_core_web_sm
    nlp = spacy.load("en_core_web_sm")
    intros = open(FILE_NAME, 'r', encoding="utf-8")
    lines = intros.readlines()
    lemma_list = []
    all_matches = []
    file_cnt = 1
    for line in lines:
        split = line.strip().split('\t')
        if len(split) > 1:
            # Get an intro.
            an_intro = line[line.index("\t"):].strip()
            # print(an_intro)
            # Pos tag it.
            doc = nlp(an_intro)
            matcher = Matcher(nlp.vocab)
            sents = list(doc.sents)
            lexv_rule = [{'POS': 'VERB'}]
            matcher.add('LexVerb', None, lexv_rule)
            matches = matcher(doc)
            for match_id, start, end in matches:
                matched_span = doc[start:end]
                # print(matched_span.text)
                all_matches.append(matched_span.text)
            # print("Processed file num:", file_cnt)
            file_cnt += 1
            if file_cnt == LIMIT:
                break
    count_frequency(all_matches)

if __name__ == "__main__":
    main()
```

Script 10: Parts of Speech (pos.py)

```
"""This script retrieves the content POS (nouns, verbs, adjectives, adverbs) found in
each article section and prints their total number."""

import spacy
from collections import Counter
nlp = spacy.load("en_core_web_sm")
f = open("abstracts.txt", "r", encoding="utf-8")
for line in f:
    if len(line) > 0:
        items = line.split("   ")
        if len(items) > 1:
            file_name = items[0]
            file_text = items[1]
            low_case = file_text.lower()
            doc = nlp(low_case)
            nouns = []
            verbs = []
            adjectives = []
            adverbs = []
            for token in doc:
                if token.pos_=="NOUN":
                    nouns.append(token.text)
                elif token.pos_ == "VERB":
                    verbs.append(token.text)
                elif token.pos_ == "ADJ":
                    adjectives.append(token.text)
                elif token.pos_ == "ADV":
                    adverbs.append(token.text)
            print(file_name)
            print("Nouns: ", nouns)
            print("Total number of nouns: ", len(nouns))
            print("Verbs: ", verbs)
            print("Total number of verbs: ", len(verbs))
            print("Adjectives: ", adjectives)
            print("Total number of adjectives: ", len(adjectives))
            print("Adverbs: ", adverbs)
            print("Total number of adverbs: ", len(adverbs))
```

Script 11: POS frequency (freq_counter.py)

```
""" The script retrieves the most frequent nouns/verbs/adjectives/adverbs per section type."""

import logging
import spacy
logging.basicConfig(level=logging.DEBUG)

# Set the section type you want to analyze.
FILE_NAME = 'abstracts.txt'
# Set this to the maximum number of files(articles) you want to process.
LIMIT = 8000
# Set this to the part of speech you want to count.
POS = 'VERB'

def count_frequency(my_list):
    # Creating an empty dictionary
    freq = {}
    for item in my_list:
        if item in freq:
            freq[item] += 1
        else:
            freq[item] = 1

    for k in sorted(freq, key=freq.get, reverse=True):
        print(k, freq[k])

def main():
    logging.info('Running freq_counter.py...')
    # Install with:
    # python -m spacy download en_core_web_sm
    nlp = spacy.load("en_core_web_sm")
    intros = open(FILE_NAME, 'r', encoding="utf-8")
    lines = intros.readlines()
    lemma_list = []
    file_cnt = 1
    for line in lines:
        split = line.strip().split('\t')
        if len(split) > 1:
            # Get an intro.
            an_intro = line[line.index("\t"):].strip()
            # print(an_intro)
            # Pos tag it.
            doc = nlp(an_intro)
            for token in doc:
                tok = token.text
                pos = token.pos_
                lem = token.lemma_
                # Check your specified part-of-speech.
                if pos == POS:
                 lemma_list.append(lem)
            print("Processed file num:", file_cnt)
            file_cnt += 1
            if file_cnt == LIMIT:
                break
    count_frequency(lemma_list)


if __name__ == "__main__":
    main()
```

Script 12: Self-Mentions (selfmentions.py)

"""The script prints the self-mentions and counts their total number per article section."""

```
import sys
import spacy
from collections import Counter
nlp = spacy.load("en_core_web_sm")
# The script below prints the self-mentions and their count in each article introduction.
# To change the section type, simply replace "introductions.txt" with
# "abstracts.txt", or
# "relatedwork.txt", or
# "discussions.txt", or
# "conclusions.txt".
f = open("introductions.txt", "r", encoding="utf-8")
#line_cnt = 0
for line in f:
   if len(line) > 0:
      items = line.split("   ")
      if len(items) > 1:
         file_name = items[0]
         file_text = items[1]
         low_case = file_text.lower()
         doc = nlp(low_case)
         Prons = ["we", "our", "us", "I", "my", "mine", "me"]
         self_mentions = []
         for token in doc:
            if token.text == token.text in Prons:
               self_mentions.append(token.text)
         print(file_name)
         print("Self-mentions found in the section:", self_mentions)
         print("Total number of self-mentions: ", len(self_mentions))
```

Script 13: Lexical Chains (lemmas.py)

```
"""The script gets all lemmas of content words and stores them in a list.
Then it iterates over the list and finds duplicates (repeated words)
as a measure for lexical cohesion. It prints the repeated word and its count (lexical
chain). It prints the total number of lexical chains. Finally, it prints the sum of all
counts."""

import spacy
from collections import Counter
nlp = spacy.load("en_core_web_sm")
f = open("relatedwork.txt", "r", encoding="utf-8")
#line_cnt = 0
for line in f:
    if len(line) > 0:
        items = line.split("    ")
        if len(items) > 1:
            file_name = items[0]
            file_text = items[1]
            low_case = file_text.lower()
            doc = nlp(low_case)
            lemmas = []
            print(file_name)
            for token in doc:
                if token.is_alpha and token.pos_ == "NOUN" or token.pos_ == "VERB" or
token.pos_ == "ADJ" or token.pos_ == "ADV":
                    lemmas.append(token.lemma_)
            def getDuplicatesfromList(listOfItems):
                dictOfElems = dict()
                for elem in listOfItems:
                    if elem in dictOfElems:
                        dictOfElems[elem] += 1
                    else:
                        dictOfElems[elem] = 1
                dictOfElems = {key:value for key, value in dictOfElems.items() if value > 1}
                return dictOfElems
            dictOfElems = getDuplicatesfromList(lemmas)
            for key,value in dictOfElems.items():
                result = key , ':', value
                # print(key , ':', value)
            print("Lexical chains: ", dictOfElems)
            count_lex_chains = len(dictOfElems)
            print("Number of lexical chains: ", count_lex_chains)
            my_values = dictOfElems.values()
            total = sum(my_values)
            print("Sum of all counts: ", total)
```

Script 14: Frequency of the repeated words (freq_lex_chains.py)

```
"""The script prints the most frequent repeated word from the lexical chains per
section type"""

import logging
import spacy
logging.basicConfig(level=logging.DEBUG)

# Setting the file containing the lexical chains per section type.
# Here it is set to the lexical chains found in the conclusion sections.
# To get the most frequent repeated word from the rest of the sections, simply replace the
FILE_NAME with :
# 'lex_chains_abstracts.txt' -- abstracts,
# 'lex_chains_intros.txt' -- introductions,
# 'lex_chains_rworks.txt' -- related works,
# 'lex_chains_discs.txt' -- conclusions.

FILE_NAME = 'lex_chains_concls.txt'
# Set this to the maximum number of files you want to process.
LIMIT = 8000

# Defining a frequency counter function
def count_frequency(my_list):
   # Creating an empty dictionary
   freq = {}
   for item in my_list:
      if item in freq:
         freq[item] += 1
      else:
         freq[item] = 1

   for k in sorted(freq, key=freq.get, reverse=True):
      print(k, freq[k])


def main():
   logging.info('Running freq_counter.py...')
   # Install with:
   # python -m spacy download en_core_web_sm
   nlp = spacy.load("en_core_web_sm")
   intros = open(FILE_NAME, 'r', encoding="utf-8")
   lines = intros.readlines()
   lemma_list = []
   file_cnt = 1
   for line in lines:
      split = line.strip().split('\t')
      if len(split) > 1:
         # Get an intro.
         an_intro = line[line.index("\t"):].strip()
```

```
            doc = nlp(an_intro)
            for token in doc:
                if token.is_alpha and token.pos_ == "NOUN" or token.pos_ == "VERB" or
token.pos_ == "ADJ" or token.pos_ == "ADV":
                    lemma_list.append(token.lemma_)
                # Check your specified part-of-speech.
            # print("Processed file num:", file_cnt)
            file_cnt += 1
            if file_cnt == LIMIT:
                break
    count_frequency(lemma_list)


if __name__ == "__main__":
    main()
```

Script 15: Connectives (connectives.py)

```
"""The script retrieves explicit connectives marking temporal, comparison,
contingency, and expansion relations and prints the total number of these connectives
per article section."""

import sys
import spacy
from collections import Counter
nlp = spacy.load("en_core_web_sm")

# The script below retrieves the connectives from each article abstract
# and prints their total number.
# To change the section type, simply replace "abstracts.txt" with
# "introductions.txt" -- for introductions,
# "relatedwork.txt" -- for related works,
# "discussions.txt" -- for discussions,
# "conclusions.txt" -- for conclusions.
f = open("abstracts.txt", "r", encoding="utf-8")
for line in f:
    if len(line) > 0:
        items = line.split("	")
        if len(items) > 1:
            file_name = items[0]
            file_text = items[1]
            low_case = file_text.lower()
            doc = nlp(low_case)
            temporal = {"after", "afterwards", "before",
                        "earlier", "later", "meanwhile",
                        "next", "previously", "simultaneously", "thereafter",
                        "till", "until", "ultimately"
                        }
            comparison = {"although", "but", "conversely", "however",
                          "instead", "nevertheless", "nonetheless", "rather",
                          "though", "whereas", "yet", "regardless", "despite", "though"
                          }
            contingency = {"as", "because", "consequently", "hence", "if",
                           "thereby", "therefore", "thus", "so", "indeed", "accordingly"}
            expansion = {"also", "alternatively", "besides", "else", "except",
                         "finally", "further", "furthermore", "likewise",
                         "moreover", "neither", "nor", "or", "otherwise", "overall", "plus",
                         "separately", "similarly", "specifically",
                         "especially", "first", "second", "firstly", "secondly"}

            temporal_relations = []
            comparison_relations = []
            contingency_relations = []
            expansion_relations = []
            for token in doc:
                if token.text == token.text in temporal:
```

```
        temporal_relations.append(token.text)
    elif token.text == token.text in comparison:
        comparison_relations.append(token.text)
    elif token.text == token.text in contingency:
        contingency_relations.append(token.text)
    elif token.text == token.text in comparison:
        expansion_relations.append(token.text)
print(file_name)
print("Temporal connectives found in the section:", temporal_relations)
print("Total number of connectives:", len(temporal_relations))
print("Comparison connectives found in the section:", comparison_relations)
print("Total number of connectives:", len(comparison_relations))
print("Contingency connectives found in the section:", contingency_relations)
print("Total number of connectives:", len(contingency_relations))
print("Expansion connectives found in the section:", expansion_relations)
print("Total number of connectives:", len(expansion_relations))
```

Prerequisites for running the scripts in the java package:

1) install Java 9 or later version
2) install OpenJDK (the java development environment)

The java scripts are also available on GitHub, as well as on GUDe under **Linguistic Mechanisms >src > sectionsretrieval.**

The java Scanner method scans the xml content line by line and looks for the section names: *Abstract, Introduction, Related Work/Background, Discussion, Conclusion/Conclusions/Future Work*. **The script below retrieves the abstracts from the research articles.**

Script 16: Section Retrieval (AbstractSection.java)

```
package sectionsretrieval;

import java.io.BufferedWriter;
import java.io.File;
import java.io.FileWriter;
import java.io.IOException;
import java.io.PrintWriter;
import java.util.ArrayList;
import java.util.Scanner;
import static java.util.stream.DoubleStream.builder;
import static java.util.stream.IntStream.builder;
import org.w3c.dom.Document;

public class AbstractRetrieval {
    public static void main(String[] args) throws IOException {
        // Accessing the directory with the xml files
        File dir = new File("../papersectionsretrieval/xmls");
        File[] files = dir.listFiles();

         // Output file
        // String path = "../sectionsretrieval/acl_anthology_sections/abstracts.txt";
        for (File file : files) {
            Scanner scan = new Scanner(file, "UTF-8");
            String contents = "";
            while (scan.hasNextLine()) {
                String myLine = scan.nextLine();
                if (myLine.startsWith("Abstract") || myLine.endsWith("Abstract") ||
myLine.endsWith("ABSTRACT")) {
                    while (!(myLine.startsWith("<sectionHeader"))&& scan.hasNextLine()) {
                        myLine = scan.nextLine();
                        // The program skips lines under the following conditions.
                        if (myLine.startsWith("<bodyText confidence") ||
myLine.startsWith("</bodyText")) {
                            continue;
                        } else if (myLine.startsWith("<table confidence") ||
myLine.startsWith("</table")) {
                            continue;
                        } else if (myLine.startsWith("<page confidence") ||
myLine.startsWith("</page")) {
                            continue;
                        } else if (myLine.startsWith("<note confidence") ||
myLine.startsWith("</note")) {
                            continue;
                        } else if (myLine.startsWith("<footnote confidence") ||
myLine.startsWith("</footnote")) {
                            continue;
                        } else if (myLine.startsWith("<figureCaption confidence") ||
myLine.startsWith("</figureCaption")) {
                            continue;
                        } else if (myLine.contains("<figureCaption confidence") ||
myLine.contains("</figureCaption>")) {
                            continue;
                        } else if (myLine.startsWith("<listItem confidence") ||
myLine.startsWith("</listItem>")) {
                            continue;
                        } else if (myLine.contains("<listItem confidence") ||
myLine.contains("</listItem>")) {
                            continue;
                        } else if (myLine.startsWith("<figure confidence") ||
myLine.startsWith("</figure>")) {
                            continue;
                        } else if (myLine.contains("<figure confidence") ||
myLine.contains("</figure>")) {
                            continue;
                        } else if (myLine.startsWith("<tableCaption confidence") ||
myLine.startsWith("</tableCaption>")) {
                            continue;
                        } else if (myLine.startsWith("<equation confidence") ||
myLine.startsWith("</equation>")) {
                            continue;
                        } else if (myLine.startsWith("<subsectionHeader confidence") ||
myLine.startsWith("</subsectionHeader>")) {
                            continue;
                        } else if (myLine.contains("<subsectionHeader confidence") ||
myLine.contains("</subsectionHeader>")) {
                            continue;
                        } else if (myLine.contains("<author confidence") ||
myLine.contains("</author>")) {
                            continue;
                        } else if (myLine.startsWith("<section confidence") ||
myLine.startsWith("</section>")) {
                            continue;
                         } else if (myLine.contains("<construct confidence") ||
myLine.startsWith("</construct>")) {
                            continue;
                        } else if (myLine.startsWith("</sectionHeader")) {
                            continue;
                        }

                        contents = contents.concat(myLine + "\n");

                    }

                    StringBuilder sb = new StringBuilder();
                    String[] lines = contents.split("\n");
                    for (String l : lines) {
                        //System.out.println("aLine: " + l);
                        if (l.endsWith("-")) {
                            sb.append(l.substring(0, l.length() - 1));

                        } else {
                            sb.append(l + " ");
                        }
                        contents = sb.toString();
                    }
                }
            }

            // System.out.println(fileID + "\t" + contents + "\n");

            String fileID = file.getName();
            FileWriter fw = new FileWriter(path, true);
            BufferedWriter bw = new BufferedWriter(fw);
            PrintWriter pw = new PrintWriter(bw);
            pw.println(fileID + "\t" + contents + "\n");
            pw.flush();
            pw.close();

        }

    }
}
```

Prerequisites for running the **Coreference.java** script

1) When you download the java package and load it in your IDE, to run the *Coreference.java* script, which retrieves the total number of coreference chains per text, make sure that the subbranch called **Libraries** is not empty (i.e. it contains Stanford CoreNLP libraries that the script requires to run the CoreNLP annotators). If the Libraries directory is empty, do the following:
2) Download Stanford CoreNLP https://stanfordnlp.github.io/CoreNLP/download.html and unarchive it.
3) Then go back to the **Libraries** subbranch, select it, and right click on it. Then select **Add JAR/Folder**, search for the corenlp package that you've already unarchived and load all libraries from the folder.

Script 17: Coreference Chains and Sentence Count (Coreference.java)

The script prints the total number of sentences, the coreference chains, and the total number of coreference chains per article section.

```
package sectionsretrieval;

import edu.stanford.nlp.coref.data.CorefChain;
import edu.stanford.nlp.ling.CoreLabel;
import edu.stanford.nlp.pipeline.CoreDocument;
import edu.stanford.nlp.pipeline.CoreSentence;
import edu.stanford.nlp.pipeline.StanfordCoreNLP;
import edu.stanford.nlp.trees.Tree;
import java.io.File;
import java.io.IOException;
import java.util.ArrayList;
import java.util.Arrays;
import java.util.List;
import java.util.Map;
import java.util.Properties;
import java.util.Scanner;
import javax.xml.parsers.DocumentBuilder;
import javax.xml.parsers.DocumentBuilderFactory;
import javax.xml.parsers.ParserConfigurationException;
import org.w3c.dom.Document;
import org.w3c.dom.Element;
import org.w3c.dom.Node;
import org.w3c.dom.NodeList;
import org.xml.sax.SAXException;


public class Coreference {

   public static void main(String... args) throws ParserConfigurationException,
SAXException, IOException {
       Properties props = new Properties();
```

```
    // set the list of annotators to run
    props.setProperty("annotators",
"tokenize,ssplit,pos,lemma,ner,parse,depparse,coref,sentiment,kbp,quote");
    // set a property for an annotator, in this case the coref annotator is being set to use the
neural algorithm
    props.setProperty("coref.algorithm", "neural");
    // build pipeline
    StanfordCoreNLP pipeline = new StanfordCoreNLP(props);

    File[] files = new File("  ").listFiles();
    Arrays.sort(files);
    analyzeFiles(files, pipeline);
  }

  public static void analyzeFiles(File[] files, StanfordCoreNLP pipeline) throws
ParserConfigurationException, SAXException, IOException {

    //Extracting the Introduction section, which can spread over several <bodyText> sections
    DocumentBuilderFactory factory = DocumentBuilderFactory.newInstance();
    DocumentBuilder builder = factory.newDocumentBuilder();
   // String path = "../sections/introductions.txt";
    for (File file : files) {
      Document doc = (Document) builder.parse(file);
      //System.out.println("Analysing file: " + file.getName());
      ArrayList<String> list = new ArrayList<>();
      // System.out.println(fileID);
      // System.out.println("------------------------------------------");
      Scanner scan = new Scanner(file, "UTF-8");
      String contents = "";
      while (scan.hasNextLine()) {
        String myLine = scan.nextLine();
        if (myLine.equals("1 Introduction") || myLine.endsWith("Introduction")) {
          while (!(myLine.startsWith("<sectionHeader"))&& scan.hasNextLine()) {
            myLine = scan.nextLine();
            // The program skips lines under the following conditions.
            if (myLine.startsWith("<bodyText confidence") ||
myLine.startsWith("</bodyText")) {
              continue;
            } else if (myLine.startsWith("<table confidence") ||
myLine.startsWith("</table")) {
              continue;
            } else if (myLine.startsWith("<page confidence") ||
myLine.startsWith("</page")) {
              continue;
            } else if (myLine.startsWith("<note confidence") ||
myLine.startsWith("</note")) {
              continue;
            } else if (myLine.startsWith("<footnote confidence") ||
myLine.startsWith("</footnote")) {
              continue;
            } else if (myLine.startsWith("<figureCaption confidence") ||
myLine.startsWith("</figureCaption")) {
              continue;
            } else if (myLine.contains("<figureCaption confidence") ||
myLine.contains("</figureCaption>")) {
              continue;
            } else if (myLine.startsWith("<listItem confidence") ||
myLine.startsWith("</listItem>")) {
              continue;
            } else if (myLine.contains("<listItem confidence") ||
myLine.contains("</listItem>")) {
              continue;
            } else if (myLine.startsWith("<figure confidence") ||
myLine.startsWith("</figure>")) {
              continue;
            } else if (myLine.contains("<figure confidence") ||
myLine.contains("</figure>")) {
              continue;
            } else if (myLine.startsWith("<tableCaption confidence") ||
myLine.startsWith("</tableCaption>")) {
              continue;
            } else if (myLine.startsWith("<equation confidence") ||
myLine.startsWith("</equation>")) {
              continue;
            } else if (myLine.startsWith("<subsectionHeader confidence") ||
myLine.startsWith("</subsectionHeader>")) {
              continue;
            } else if (myLine.contains("<subsectionHeader confidence") ||
myLine.contains("</subsectionHeader>")) {
              continue;
            } else if (myLine.contains("<sectionHeader confidence") ||
myLine.contains("</sectionHeader>")) {
              continue;
            } else if (myLine.startsWith("</sectionHeader")) {
              continue;
            }

            contents = contents.concat(myLine + "\n");

          }

          StringBuilder sb = new StringBuilder();
          String[] lines = contents.split("\n");
          for (String l : lines) {
            //System.out.println("aLine: " + l);
            if (l.endsWith("-")) {
              sb.append(l.substring(0, l.length() - 1));

            } else {
              sb.append(l + " ");
            }
            contents = sb.toString();
          }

        }
      }
      CoreDocument document = new CoreDocument(contents);
      pipeline.annotate(document);
      List<CoreSentence> sentencesOfDoc;
      sentencesOfDoc = document.sentences();
      // Total number of sentences
      int sentCount = document.sentences().size();
      // Getting the corefchains in the document
      Map<Integer, CorefChain> corefChains = document.corefChains();
      int corefNum = corefChains.size();
      //System.out.println(contents);
      String fileID = file.getName();
      System.out.println(fileID);
     // System.out.println(document);
      System.out.println(sentCount);
      System.out.println(corefChains);
      System.out.println(corefNum);
    }

  }

}
```

Bibliography

Adorján, Maria. 2013. *Lexical Repetition in Academic Discourse. A Computer-Aided Study of the Text-organizing Role of Repetition.* https://www.grin.com/document/459913 (28.05.2021).

Ahmad, Jameel. 2012. *Stylistic Features of Scientific English: A Study of Scientific Research Articles*. In English Language and Literature Studies, 2(1), 47-55. https://doi.org/10.5539/ells.v2n1p47 (04.07.2020).

Austin, John. 1962. *How to Do Things with Words.* Cambridge, MA: Harvard University Press.

Benz, Anton & Katja Jasinskaja. 2017. *Questions under discussion: From sentence to discourse*. In Discourse Processes, 54(3), 177-186. https://doi.org/10.1080/0163853X.2017.1316038 (04.07.2020).

Brown, Gillian & George Yule. 1983. *Discourse Analysis.* Cambridge: Cambridge University Press.

Büring, Daniel. 2003. *On D-trees, beans and B-accents.* In Linguistics and Philosophy, 26(5), 511–545.

Chafe, Wallace. 1976. *Givenness, contrastiveness, definiteness, subjects, topics, and point of view.* In Subject and Topic. New York: Academic Press, 25-36.

Cohen, Lisa M. 2020. *The new era of disinformation wars.* Völkerrechtsblog. http://nbn-resolving.de/urn:nbn:de:0301-20210107-183308-0-1 (04.06.2021).

Devlin, Jacob, Ming-Wei Chang, Kenton Lee & Kristina Toutanova. 2018. *Bert: Pre-training of deep bidirectional transformers for language understanding*. https://arxiv.org/abs/1810.04805 (04.07.2020).

Enkvist, Nils E. 1990. *Seven problems in the study of coherence and interpretability.* In Connor, U. and Johns, A. (eds.) Coherence in writing: Research and pedagogical perspectives. Washington, DC: TESOL Publications, 9-28.

Fasold, Ralph W. 1990. *Sociolinguistics of Language*. Oxford: Blackwell.

Flower, Linda & John R. Hayes. 1981. *A cognitive process theory of writing*. In College Composition and Communication, 32(4), 365-387.

Foucault, Michel. 1982. *Power/Knowledge: Selected Interviews and Other Writings by Michel Foucault, 1972-1977*. New York: Pantheon.

Furkó, Péter B. 2020. *Discourse Markers and Beyond*. Palgrave Macmillan.

Ginzburg, Jonathan. 2012. *The Interactive Stance: Meaning for Conversation*. Oxford: Oxford Univ. Press.

Grimshaw, Allen. 1981. *Language as Social Resource*. Stanford, CA: Stanford University Press.

Goffman, Erving. 1963. *Behaviour in Public Places*. New York: Free Press.

Grosz, Barbara J., Aravind K. Joshi & Scott Weinstein. 1995. *Centering: A framework for modeling the local coherence of discourse.* In Computational Linguistics, 21(2), 203-225.

Grosz, Barbara J. & Candace Sidner. 1986. *Attention, Intentions, and the Structure of Discourse.* Computational Linguistics, 12(3), 175-204.

Gumperz, John. 1982. *Discourse Strategies.* Cambridge: Cambridge University Press.

Halliday, Michael & Ruqaiya Hasan. 1976. *Cohesion in English*. Longman Group Ltd London.

Harris, Zellig .1952. *Discourse Analysis*. In Language, 28(1), 1-30.

Harris, Zellig. 1988. *Language and Information*. New York: Columbia University Press.

Hoey, Michael. 1991. *Patterns of lexis in text*. Oxford: Oxford University Press.

Hübler, Axel. 1986. *Understatements and Hedges in English.* Amsterdam/Philadelphia: John Benjamins Publishing Company.

Hyland, Ken. 1998. *Hedging in Research Articles*. John Benjamins Pub. Co. https://doi.org/10.1075/pbns.54 (01.06.2021).

Hyland, Ken. 2001. *Humble servants of the discipline? Self-mention in research articles*. English for Specific Purposes Volume 20(3), 207-226. https://doi.org/10.1016/S0889-4906(00)00012-0 (04.03.2021).

Ivanova, Iverina. 2020. *Assessing the Effect of Text Type on the Choice of Linguistic Mechanisms in Scientific Publications*. In Proceedings of the ESSLLI & WeSSLLI Student Session 2020, 25-35. https://www.brandeis.edu/nasslli2020/pdfs/student-session-proceedings-compressed.pdf (01.06.2021).

Keskar, Nintish S., Bryan McCann, Lav R. Varshney, Caiming Xiong & Richard Socher. 2019. *CTRL: A conditional transformer language model for controllable generation*. CoRR. https://arxiv.org/abs/1909.05858 (04.07.2020).

Lakoff, George. 1973. *Hedges: A study in meaning criteria and the logic of fuzzy concepts*. Chicago Linguistic Society Papers, 8, 183-228. https://escholarship.org/uc/item/0x0010nv (15.01.2021).

Linde, Charlotte & William Labov. 1975. *Spatial networks as a site for the study of language and thought*. In Language, 51(4), 924-939. https://www.jstor.org/stable/412701?seq=1#metadata_info_tab_contents (04.02.2021).

Lu Chao, Yi Bu, Jie Wang, Ying Ding, Vetle Torvik, Matthew Schnaars & Chengzhi Zhang. 2018. *Examining Scientific Writing Styles from the Perspective of Linguistic Complexity.* https://doi.org/10.1002/asi.24126 (28.05.2021).

Mahyoob, Mohammad, Jeehaan Algaraady & Musaad Alrahaili. 2020. *Linguistic-Based Detection of Fake News in Social Media.* In International Journal in English Linguistics, 11(1) 2021, 99-109. https://doi.org/10.5539/ijel.v11n1p99 (04.06.2021).

Mann, William C. & Sandra A. Thompson. 1988. *Rhetorical Structure Theory: Toward a functional theory of text organization*. In Text, 8(3), 243–281. https://doi.org/10.1515/text.1.1988.8.3.243 (25.11.2020).

Myers, Greg. 1991. *Writing Biology: Texts in the Social Construction of Scientific Knowledge.* In Journal of the History of Biology, 24(3), 521-527.

McNamara, Danielle S., Scott A. Crossley & Philip M. McCarthy. 2009. *Linguistic Features of Writing Quality*. In Written Communication, 27(1), 57-86. https://doi.org/10.1177/0741088309351547 (02.05.2021).

McNamara, Danielle S. & Arthur C. Graesser. 2011. *Coh-Metrix: An Automated Tool for Theoretical and Applied Natural Language Processing.* In P. McCarthy, C. Boonthum-Denecke (eds.), Applied Natural Language Processing: Identification, Investigation and Resolution, 188-205, Hershey, PA: IGI Global.

Orasan, Constantin. 2001. *Patterns in Scientific Abstracts*. In Proceedings Corpus Linguistics, 433-445.

Pitler, Emily & Ani Nenkova. 2009. *Using Syntax to Disambiguate Explicit Discourse Connectives in Text.* In Proceedings of the {ACL}-{IJCNLP} 2009 Conference Short Papers, Association for Computational Linguistics, 13-16. https://www.aclweb.org/anthology/P09-2004https://www.aclweb.org/anthology/P09-2004/ (28.05.2021).

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei & Ilya Sutskever. 2019. *Language models are unsupervised multitask learners*. https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf (04.07.2020).

Rey, Denise & Markus Neuhauser. 2011. *Wilcoxon-Signed-Rank Test*. International Encyclopedia of Statistical Science. Heidelberg Springer, 1658 - 1659.

Roberts, Craige. 1996. *Information structure in discourse: Towards an integrated formal theory of pragmatics.* OSU Working Papers in Linguistics, 49, 91–136.

Sarkar, Dipanjan. 2019. *Text Analytics with Python*. Apress, Berkeley, CA.

Sarts Janis. 2021. *Disinformation as a Threat to National Security*. In Jayakumar S., Ang B., Anwar N.D. (eds) Disinformation and Fake News. Palgrave Macmillan, Singapore. https://doi.org/10.1007/978-981-15-5876-4_2 (04.06.2021).

Schiffrin, Deborah. 1994. *Approaches to Discourse*. Oxford: Blackwell.

Schleppegrell, Mary J. 2004. *The language of schooling; a functional linguistics perspective*. London: Lawrence Erlbaum Associates. http://www.iltec.pt/TeL4ELE/Schleppegrell.pdf (12.10.2020).

Stalnaker, Robert. 1978. *Assertion*. In Peter Cole (ed.), Syntax and Semantics, 9, Pragmatics. New York Academic Press, 315-332.

Stalnaker, Robert. 2002. *Common Ground*. In Linguistics and Philosophy, 25, 701–721. https://doi.org/10.1023/A:1020867916902 (28.05.2021).

Tompkins, Jillian. 2019. *Disinformation Detection: A review of linguistic feature selection and classification models in news veracity assessments.* https://arxiv.org/abs/1910.12073v1 (04.06.2021).

van Kuppevelt, Jan. 1995. *Discourse structure, topicality and questioning*. In Journal of Linguistics, 31(1), 109–147.

Vanlangendonck, Flora, Roel Willems, Laura Menenti & Peter Hagoort. 2013. *The role of common ground in audience design: Beyond an all or nothing story.* https://pre2013.uvt.nl/pdf/vanlangendonck-willems-menenti-hagoort.pdf (28.05.2021).

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser & Illia Polosukhin. 2017. *Attention is all you need.* In I. Guyon, et al.,editors, Advances in Neural Information Processing Systems 30, 5998-6008. Curran Associates, Inc. https://papers.nips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf (22.06.2021).

Veel, R. 1997. *Learning how to mean-scientifically speaking: Apprenticeship into scientific discourse in the secondary school.* In F. Christie & J. R. Martin (eds.), Genre and institutions: Social processes in the workplace and school, London: Cassell, 161–195.

von Stutterheim, Christiane & Wolfgang Klein. 1989. *Referential Movement in Descriptive and Narrative Discourse.* North-Holland Linguistic Series: Linguistic Variations, Elsevier, 54, 39-76. https://www.sciencedirect.com/science/article/pii/B9780444871442500057?via%3Dihub (23.1.2021).

Wang Jiayu & Yi Zhang. 2019. *Lexical Cohesion in Research Articles*. In Linguistics and Literature Studies, 7(1), 1 - 12. https://www.hrpub.org/download/20181230/LLS1-19312453.pdf (22.06.2021).

Webber, Bonnie, Rashmi Prasad, Eleni Miltsakaki, Nikhil Dinesh, Alan Lee, Aravind Joshi & Livio Robaldo. 2007. *The Penn Discourse Treebank 2.0 Annotation Manual*, 8-21. https://www.seas.upenn.edu/~pdtb/PDTBAPI/pdtb-annotation-manual.pdf (28.05.2021).

Wellington, Jerry & Jonathan Osborne. 2001. *Language and literacy in science education.* Buckingham: Open University Press.

Witte, Stephen P. & Lester Faigley. 1981. *Coherence, cohesion, and writing quality*. College Composition and Communication, 32(2), 189-204.

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz & Jamie Brew. 2019. *HuggingFace's transformers: State-of-the-art natural language processing*. https://arxiv.org/abs/1910.03771v4 (04.07.2020).

Yazilarda, Akademik, Yazari İşaret, Eden S. Kullanımı & Huseyin Kafes. 2017. *The use of authorial self-mention words in academic writing*. International Journal of Language Academy, 5(3), 165-180. https://oaji.net/articles/2017/505-1499642655.pdf (04.07.2020).