Funded by the European Union

UK Research and Innovation

Deliverable number: D1.2

hybrids

Technical report on the state of the art on NLP applied to discourse analysis

Hybrid Intelligence approaches to computational discourse studies.

Version 1.

Project Details

Project Acronym:	HYBRIDS
Project Title:	Hybrid Intelligence to monitor, promote and analyse transformations in good democracy practices
Grant Number:	101073351
Call	HORIZON-MSCA-2021-DN-01
Topic:	HORIZON-MSCA-2021-DN-01-01
Type of Action:	HORIZON-TMA-MSCA-DN
Project website:	https://hybridsproject.eu/
Coordinator	Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS)-Universidade de Santiago de Compostela (USC)
Main scientific representative:	Prof. Pablo Gamallo Otero, pablo.gamallo@usc.es
E-mail:	citius.kmt@usc.es, info@hybridsproject.eu
Phone:	+34 881 816 414

Deliverables Details

Number:	D1.2
Title:	Technical report on the state of the art on NLP applied to discourse analysis
Work Package	WP1: Hybrid Intelligence
Lead beneficiary:	UDC
Deliverable nature:	R- Document, report
Dissemination level:	PU-Public
Due Date (month):	30/06/2024 (M18)
Submission Date (month):	30/06/2024 (M18)
Keywords:	Natural Language Processing, Discourse, Argumentation, Dialogue,

Abstract

D1.2. Technical report on the state of the art on NLP applied to discourse analysis offers a systematic review of scientific approaches within the computational treatment of discourse, both at a formal, methodological, and algorithmic level. The report covers in a transversal way two main aspects of the relationship between discourse and its technological approaches. Firstly, the main formal theories for representing discourse computationally. Secondly, the report addresses existing algorithms and methods for several computational tasks related to discourse in a multilingual sphere. Thus, this document represents the theoretical discourse formalisms and the up-to-date algorithms that will inform the implementation of new tools and approaches for a hybrid intelligence discourse treatment.

Deliverable Contributors

	Name	Institution	E-mail
Deliverable leader	Patricia Martín Rodilla	UDC	patricia.martin.rodilla@udc.es
Contributing Authors	Martial Pastor	RU	martial.pastor@ru.nl
	Davide Bassi	CiTIUS-USC	davide.bassi@usc.es
	Siddharth Bhargava	FBK	sbhargava@fbk.eu
	Erik Marino	UEVORA	erik.marino@uevora.pt
	Katarina Laken	FBK	alaken@fbk.eu

History of Changes

Version	Date	Changes to previous version	Status
0.1	07/06/2024	First Draft	Draft
0.2	27/06/2024	Consortium Internal review	Review
1.0	27/06/2024	Approved version to be submitted	Final

List of Acronyms

AIF	Argument Interchange Framework
CISA	Classroom Interaction Speech Act
EDU	Elementary Discourse Unit
IAT	Inference Anchoring Theory
LDA	Latent Dirichlet analysis
LIWC	Linguistic Inquiry and Word Count
LLM	Large Language Model
ML	Machine Learning
NLP	Natural Language Processing
PDTB	Penn Discourse Treebank
PLTM	Polylingual Topic Model
RST	Rhetorical Structure Theory
SAT	Speech Acts Theory
SOTA	State-of-the-art
SVM	Support Vector Machine

Contents

1 Introduction **6**

2 Discourse Parsing and Rhetorical Structure Theory **7**
2.1 Introduction: Discourse Parsing . . . 7
2.2 Rhetorical Structure Theory . . . 7
2.2.1 Discourse Structure and RST-DT . . . 8
2.2.2 RST-DT: The NLP task . . . 8
2.2.3 RST Discourse Parsing: Parser Architecture and Performance . . . 9
2.2.4 RST Signals and Error Analysis . . . 11
2.3 Conclusion . . . 12

3 An Argumentative Perspective on the Dialogic Discourse **13**
3.1 Introduction to Dialogical Argumentation . . . 13
3.2 Modeling Dialogical Argumentation . . . 17
3.3 Symbolic Approaches to Dialogical Argumentation . . . 20
3.4 Computational Approaches to Dialogical Argumentation . . . 22
3.5 Conclusion . . . 23

4 From Words to Functions: Analyzing Discourse through "Speech Acts Theory" **25**
4.1 Speech Acts Theory from a Philosophy of Language Perspective . . . 25
4.2 NLP models for Automatic Speech Acts Classification . . . 27
4.2.1 Offline Interactions Analysis . . . 27
4.2.2 Online Interaction Analysis . . . 28
4.2.3 Abusive Language . . . 29
4.2.4 Political Analysis . . . 30
4.2.5 News Analysis . . . 31
4.2.6 Health Promoting Communication . . . 33
4.3 Conclusions . . . 33
4.3.1 Limitations and Future Directions . . . 34

5 Discourse Analysis Detection of Conspiracy Theories Using NLP Computational Techniques **36**
5.1 Introduction to Conspiracy Theories in Discourse Analysis . . . 36
5.2 History of Combating Misinformation: **Overview of computational techniques to detect general misinformation** . . . 36
5.3 Challenges in Analyzing Conspiracy Theories with NLP . . . 38
5.4 Computational Approaches Specific to Conspiracy Theories . . . 39
5.5 Future Directions and Research Opportunities . . . 41
5.6 Conclusion . . . 42

6 Multilingual approaches to computational discourse analysis 43
6.1 Multilingual NLP for rhetorical analysis 43
6.2 Multilingual approaches to measuring discourse coherence 45
6.3 Multilingual approaches to topic modeling 46
6.3.1 Evaluation 50
6.3.2 Challenges and future work 51
6.4 Conclusion 51

7 Overall Conclusion 51

1 Introduction

This technical report aims to provide a deep perspective on the theoretical, methodological, and technological foundations currently supporting discourse technological treatment, understanding discourse as ”the use of language in context” Fairclough (1996). This is an internal report of great value within the objective of the HYBRIDS project, since it tries to identify unequivocally the different contributions made to date that support a true hybrid approach between the philological, social, philosophical, and communication approaches to the discourse concept, and the current algorithms developed for its identification, treatment, analysis, detection, and prediction.

In this context, this document is structured into five sections that allow us to address this discourse-technological intersectionality from different angles. Firstly, section 2 fulfills a double objective: it formally presents the Rhetorical Structure Theory (RST) discourse theory, and taking advantage of its possibilities in terms of computational formalization, it addresses the specific problem of the automatic discourse parsing task. Given a discursive fragment, Can we currently obtain a discursive scheme automatically? It details RST as a formalism to express discourse and the current algorithms with the best performance in discourse parsing task.

Linking to the more computational approaches, section 2 has numerous connections with Section 6, where multilingual approaches in natural language and speech processing are addressed. Most multilingual approaches express discourse using RST to later use advanced language models (Large Language models, LLMs) to provide algorithms with multilingual capabilities.

In addition, the current trends in discourse and technology are not only formalized under RST, but there are numerous alternative approaches whose degree of computation had been much lower until a few years ago, but whose current works present great potential in computational tasks. Thus, this document also addresses 1) purely dialogue-based approaches in Section 3 and their possible computation alternatives 2) the formal framework Speech Act Theory in Section 4, with an enormous current development in computational supports and developments in tasks of special relevance for HYBRIDS, such as detection of offensive content or inaccurate news and 3) specific formalisms for conspiracy theories that allow exemplifying the discursive treatment without previous formal theories presented in Section 5.

2 Discourse Parsing and Rhetorical Structure Theory

2.1 Introduction: Discourse Parsing

Discourse parsing has sparked significant interest in recent NLP applications. This task goes beyond the conventional scope of sentences and may extend to encompass the identification of Coherence Relations (relations between segments of text)at the discourse level.

Among the various computational approaches to discourse, we recognize Rhetorical Structure Theory (RST) style (Carlson et al. (2001)and PDTB style discourse parsing (Stede and Neumann (2014). An RST-style discourse parser aims to derive a hierarchical rhetorical tree structure from an input document, whereas a PDTB discourse parser seeks to establish a flat discourse structure between sentences or clauses rather than a tree. Additionally, we note that discourse parsing for multiparty dialogues generates a discourse dependency graph from the input dialogue. Unlike RST and PDTB styles, this approach allows discourse relations to exist between non-adjacent utterances.

In this section we focus on one of the most popular formalisms for representing discourse : Rhetorical Structure Theory Mann and Thompson (1987), which has spurred the construction of various datasets that are now used for hierarchical discourse parsing. This last task is challenging and discourse parsers have not achieved the same level of success as other tasks at the sentence level. Moreover, analyzing failure cases, especially in deep learning-oriented parsers, proves difficult.

We first outline the framework based on classical RST theory Mann and Thompson (1987) and annotation guidelines Stede et al. (2017). Next, we describe the RST discourse parsing task and discuss how existing corpora have been used for training and evaluating discourse parsers. We then explore various advancements in parser architectures and their performance, from the initial models to state-of-the-art architectures, while also examining how error analysis has been conducted and the conclusions drawn from it. Finally, we conclude by citing notable works and applications in discourse parsing and discourse features, and offer perspectives for future research.

2.2 Rhetorical Structure Theory

Rhetorical Structure Theory (RST) was originally developed for text generation and creating computer programs with some author-like capabilities. It has since become fundamental to many text-based applications, including question-answering (Chai and Jin (2004) and dialogue generation (Prendinger et al. (2007)). With the rise of Deep Learning in NLP, discourse structures have gained significant interest in two main areas: predicting discourse structures has become a popular task in itself (Zeldes et al. (2021)), and discourse features are also considered important for various NLP tasks involving the analysis of persuasion or argumentation strategies (Chernyavskiy et al. (2024) ; Pastor et al. (2024)).

2.2.1 Discourse Structure and RST-DT

RST is a model for textual analysis of coherence relations. With this model, the relations between text segments are annotated with different classes of coherence relations such as elaboration, contrast, causal, temporal, etc. The text segments in an RST tree are ”elementary discourse units” (EDUs), which are contiguous sets of tokens approximately similar to independent clauses. The relations occur not only between text segments but also between groups of text segments, which means that the final RST representation of a complete text (book, chapter, article, comment, etc.) is a hierarchical tree of text segments connected by coherence relations as shown in Figure 1.

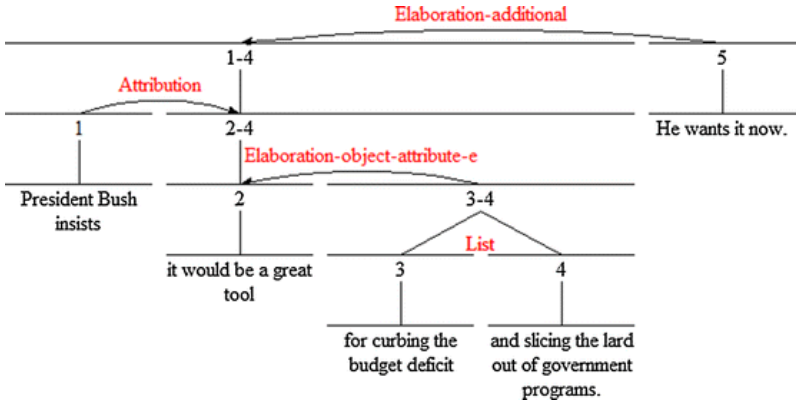


Figure 1: RST analysis from document wsj_609 in the RST-DT corpus, which describes the textual analysis model of coherence relations.

Above, we can notice arrows pointing to particular structures or EDUs. These arrows differentiate between the nucleus and the satellites, a differentiation that operates for each mononuclear relation in the tree. The direction of the arrow points towards the nucleus. The differentiation between the nucleus and the satellite is based on the principle that a text unit acting as a satellite can be removed without altering the coherence of the discourse, whereas removing a text unit acting as a nucleus would render the text incoherent.

2.2.2 RST-DT: The NLP task

RST-style discourse parsing involves two main tasks: discourse segmentation and tree construction. The segmentation module divides the input text into elementary discourse units (EDUs), while the tree construction module creates an RST tree structure using these EDUs. Given that the EDU segmentation module has achieved a performance rate nearing 95%, most research in RST-style discourse parsing now concentrates on the RST tree construction task with predefined gold EDUs (Li et al. (2022a)).

Discourse segmentation. The EDU segmentation task aims to segment input text into elementary discourse units (EDUs). EDUs are the minimal discourse unit in RST-style discourse parsing, and they are typically clauses. Annotators label EDUs according to the following rules:

1. Clauses that serve as subjects or objects of the main verb are not considered EDUs.

2. Clauses that function as complements of the main verb are not considered EDUs.
3. Complements of attribute verbs (e.g., speech acts and other cognitive acts) are considered EDUs.
4. Relative clauses, nominal postmodifiers, or clauses that interrupt other valid EDUs are treated as embedded discourse units.
5. Phrases that start with a strong discourse marker, such as *because*, are treated as EDUs.

RST tree building. For the RST tree building task, golden EDUs are already provided and the task aims to construct an RST tree and label rhetorical relations on links. In detail, this task contains the following subtasks: span prediction, nuclearity indication, and relation classification.

- **Span prediction**: This subtask can be regarded as a binary classification task that aims to predict the tree structure of input text by classifying whether two EDUs or spans should be merged.
- **Nuclearity indication**: As mentioned above, there are two different kinds of nodes in the RST tree for hypotactic relations: nucleus and satellite. The nuclearity indication task aims to predict the nucleus or satellite given two EDUs or spans.
- **Relation classification**: This subtask aims to classify the specific rhetorical relations between given two EDUs or spans. In RST-DT, there are 78 fine-grained rhetorical relations in total, including 53 mononuclear relations and 23 multi-nuclear relations. The definitions of a specific rhetorical relation are based on constraints on the nucleus, constraints on the satellite, constraints on the combination of nucleus and satellite, and effect achieved on the text receiver.

2.2.3 RST Discourse Parsing: Parser Architecture and Performance

The task of discourse parsing has evolved significantly, aligning with advancements in computational linguistics and NLP techniques. In this overview, we will highlight the most prominent efforts to develop discourse parsers for RST. We will begin by examining rule-based methods, followed by feature-based machine learning approaches, and conclude with state-of-the-art parsers utilizing deep learning algorithms. On Figure 2 we can see the F1-score metric performance for span prediction.

Rule based methods. Rules are usually based on automatically derived sentential syntactic structures, CPs, constraints related to textual adjacency and organization, and cohesive constructs Hou et al. (2020). The text-level parser incorporates constraints on textual adjacency and organization into a beam search to generate optimal RS-trees.

In discourse segmentation, Tofiloski et al.,2009 argued that rule-based methods have certain advantages over machine learning methods in avoiding the over-segmentation problem. Their proposed discourse segmenter, SLSeg, utilizes syntax-based rules to determine EDU boundaries. After analyzing the characteristics of EDUs, the authors assumed that each EDU contains a verb. SLSeg includes 12 syntax-based rules and a few lexical rules involving a list of stop phrases, CPs, and word-level POS tags. SLSeg was tested on nine manually annotated texts

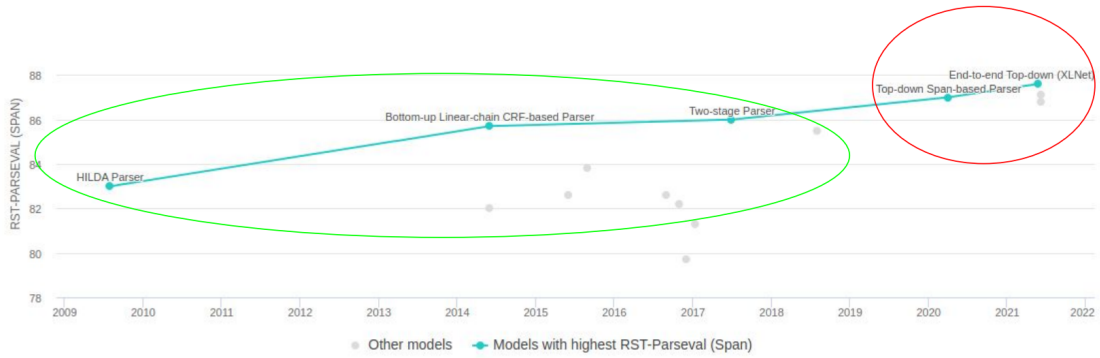


Figure 2: F1-score Evolution of discourse parsing performance, from papers with code, for span prediction. Circled in Green are feature based machine learning methods while Red are deep learning based.

with an inter-annotation agreement of 85%, including three texts from the RST website. Experimental results indicated that SLSeg achieved higher precision compared to other elemenatary feature based machine learning. However, this method was evaluated on a small corpus, and the number of rules required to accommodate diverse written texts would be very large.

Rules can detect EDU boundaries and determine rhetorical relations between text spans. However, for heterogeneous and lengthy texts, a vast number of rules are needed, leading to combinatorial explosion problems, making the approach time-consuming and costly.

Feature based machine learning. In the field of machine learning and discourse parsing, we should mention Sagae, 2009 who proposed a model comprising two main procedures: discourse segmentation and discourse parsing. Discourse segmentation is modeled as a binary classifier that determines whether to insert an EDU boundary after a word, utilizing features such as syntactic dependency labels and the POS tags of the current and neighboring words. This classification is performed using an averaged perceptron. The transition-based discourse parsing involves a straightforward percolation of ”head EDUs,” resulting in significant improvements in accuracy and speed on the RST-DT corpus compared to previous methods. Additionally, this algorithm is highly efficient, operating in linear time.

We should also mention the HILDA parser as a noteworthy effort in this field. HILDA is a document-level RST parser developed by Hernault et al., 2010. The first step in HILDA's process is discourse segmentation, where a combination of syntactic and lexical features is used to train a binary classifier to segment the text into EDUs. Subsequently, two SVM classifiers are employed to construct the tree structure in a greedy manner. The resulting RS-tree is always a binary tree, with multi-nuclear rhetorical relations binarized into a right-branching binary tree. The first SVM classifier determines the probability of a structural relation between consecutive EDUs, merging the pair with the highest probability. The second classifier, a multi-class SVM, selects rhetorical relations for the merged EDUs. This process is repeated until the full RS-tree is built. The features used for relation labeling include several shallow lexical and syntactic features. For the CP, instead of using a pre-defined CP dictionary, an empirical 3-gram dictionary from the training

corpus is constructed. HILDA was evaluated on the RST-DT corpus using 18 high-level rhetorical relations and is noted as the first fully implemented text-level RST parser.

Deep Learning based approaches. Deep learning architectures have demonstrated the ability to capture relevant representations that conventional machine learning-based methods often miss. In RST parsing, the main idea approach is to transform each discourse unit into an abstract vector representation. These representations are then used at later stages of the parsing process to compute the most likely structures and to classify relations between these same structures or between spans of text. EDU abstract vector representations can be used differently, at various stages of the parsing process. Here, we limit the presentation to parsers that achieved state-of-the-art performance before 2024.

For instance, in a top-down parsing approach, Liu et al., 2021a propose document-level neural parser includes several key components. Initially, a segmenter predicts the EDU breaks, followed by a hierarchical encoder that generates the EDU representations. Then, a pointer-network-based decoder and a relation classifier work together to predict the tree structure, nuclearity, and rhetorical relations. The decoder maintains a stack to track top-down, depth-first span splitting. For each splitting point, sub-spans are fed to a classifier to determine nuclearity and relations.

In a bottom-up approach, Nguyen et al. (2021),2021 consider discourse parsing as a sequence of splitting decisions at token boundaries, using a seq2seq network to model these decisions. Their framework enables discourse parsing from scratch without requiring prior discourse segmentation; instead, it produces segmentation as part of the parsing process. The unified parsing model employs a beam search to decode the optimal tree structure by exploring a space of high-scoring trees. Through extensive experiments on the standard English RST discourse treebank, the authors demonstrate that their parser significantly outperforms existing methods in both end-to-end parsing and parsing with gold segmentation. Moreover, it achieves this without using any handcrafted features, making it faster and more easily adaptable to new languages and domains.

2.2.4 RST Signals and Error Analysis

Early research on discourse parsing has highlighted the importance of discourse markers (DMs). Studies, demonstrate that DMs like ”if” can make certain relations, such as CONDITION, easier to identify (Pitler et al., 2008). This focus is particularly evident in shallow discourse parsing using the Penn Discourse Treebank (PDTB), which concentrates on local text relations. Explicit relations, often signaled solely by DMs, are considered the easiest to recognize. Neural parsers using contextualized embeddings have shown high efficacy in identifying these explicit relations, achieving notable F1 scores for both explicit and implicit relations.

In contrast, investigations into DMs within the RST framework, such as those by Das and Taboada, 2019, are less common. Liu et al., 2023 specifically explored the role of DMs in RST parsing, finding that while DMs are influential, intra-sentential features can be more significant in predicting relation labels. This suggests that explicit relations are not exclusively signaled by DMs but also by other textual elements, indicating the need for further research into these additional signals of coherence relations.

In the context of these findings, Pastor and Oostdijk, 2024 have presented an approach for

assessing the importance of Das and Taboada, 2018 signals within the context of discourse parsing. Their initial observations reveal distinct patterns in the performance of a discourse parser when graphed for specific signals, leading to various implications. Initially, it is noted that DMs are not consistently reliable signals for all relationships; in fact, they can be viewed as distractors, causing confusion between relations signaled by the same DMs. Subsequently, an examination of the effectiveness of alternative signal types, including syntactic, semantic, and genre-related signals, is conducted. The findings demonstrate that, despite certain syntactic signals not being predominant for specific relations, they still prove to be effective. Subsequently, the authors conduct an experiment incorporating the modeling of RST signals as features for a parser error or parser success prediction model. The results demonstrate the relevance of utilizing signals as features, providing valuable insights into the signals (or combination of signals) that facilitate relation recognition. Moreover, their observations also shed light on scenarios where the presence of specific signals might pose challenges or lead to confusion, making it difficult for the parser to accurately discern certain relations.

2.3 Conclusion

The rise of Deep Learning in NLP has significantly increased interest in discourse structures, both as a task in itself and as a crucial component for various NLP tasks such as analyzing persuasion or argumentation strategies (Zeldes et al., 2021; Chernyavskiy et al.,2024; Pastor et al.,2024) The segmentation module, responsible for dividing text into elementary discourse units (EDUs), has achieved a performance rate nearing 95%. Consequently, current research predominantly focuses on the RST tree construction task using predefined gold EDUs.

Discourse parsing has evolved alongside developments in computational linguistics and NLP techniques. Various parser architectures have been proposed, ranging from rule-based methods to deep learning algorithms. Rule-based methods utilize predefined rules to detect EDU boundaries and determine rhetorical relations between text spans. However, these methods are limited by their scalability and complexity, particularly when dealing with heterogeneous and lengthy texts, leading to combinatorial explosion problems.

Feature-based machine learning approaches, such as HILDA, have demonstrated significant advancements by incorporating high-level rhetorical relations and performing evaluations on the RST-DT corpus. These methods marked a transition towards more sophisticated models capable of handling text-level RST parsing comprehensively.

Deep learning architectures have further advanced RST parsing by capturing relevant representations that traditional machine learning methods often miss. The main approach involves transforming each discourse unit into an abstract vector representation, which is then used in subsequent stages to compute likely structures and classify relations between spans of text. These state-of-the-art parsers have shown remarkable performance improvements, establishing new benchmarks in the field.

Error analysis in RST discourse parsing has revealed crucial insights into the effectiveness of various signals. Pastor and Oostdijk, 2024 assessed the importance of signals such as discourse markers (DMs) and found that they are not consistently reliable for all relationships, often causing confusion. Alternative signals, including syntactic, semantic, and genre-related cues, were

examined, showing that despite some syntactic signals not being predominant for specific relations, they still prove effective. Modeling RST signals as features for predicting parser errors or successes highlighted the relevance of using these signals to facilitate relation recognition. This analysis also identified scenarios where certain signals could pose challenges, making accurate relation discernment difficult for parsers.

In summary, the field of RST discourse parsing has progressed significantly with the advent of deep learning, moving from rule-based methods to sophisticated deep learning architectures. Ongoing error analysis continues to refine our understanding of effective signal utilization, ultimately contributing to the development of more robust and accurate discourse parsers.

3 An Argumentative Perspective on the Dialogic Discourse

This section focuses on providing an overview on dialogical argumentation, with a strong focus on presenting Natural Language Processing (NLP) based techniques and tools that study two of the main dialogue types, persuasion and deliberation. As conversational AI (AI personal assistants/agents and Bot-as-a-Service (BaaS)) takes the forefront role in recent technological advancements of the world, it is imperative to shed light on the various discourse and computational software and tools that are available for analyzing conversations in different contexts and settings, whether from a symbolic perspective or a computational one. Likewise, Social media has cemented itself as the leading platform for acquiring, exchanging and propagating new information at a scale, unprecedented in terms of volume, velocity and variety. With new Generative AI and LLMs gaining momentum, we see Conversational AI observing exponential growth in development and research across industries requiring human-computer interaction. This chapter will highlight how argumentation can play an important role in online discussion and debate platforms for improving critical reasoning and learning capabilities of the users, building efficient decision making models capable of evidence-based practical argumentation, and understanding how information exchanges occur on the platforms.

3.1 Introduction to Dialogical Argumentation

Dialogical argumentation finds itself in a unique intersection of three main and emerging technological advancements, namely argumentation modeling, conversational artificial intelligence, and social media analysis. Argumentation models have become increasingly important in computer science and artificial intelligence (Lippi and Torroni, 2016; Macagno, 2021; Lytos et al., 2019), finding application across multiple industries and sciences such as biology, genetics, political science, and law. Social media has become the primary source for opinion formation, information exchange, and mass communication (Sapountzi and Psannis, 2018; Álvarez-Peralta et al., 2023). Consequently, there has been an increase in information manipulation, opinion polarization, and harmful content propagation, leading to a need for effective regulation and information moderation (Silva, 2016; Myers West, 2018; Wilson and Land, 2021; Gearhart et al., 2020). Finally, as conversational artificial agents, i.e. chatbots such as ChatGPT (https://openai.com/index/chatgpt/) and Gemini (https://gemini.google.com/app), start becoming the main tools for information search and

knowledge generation, it is imperative to build tools that allow for efficient human-machine communication and understanding (Bansal et al., 2024). The focus on building efficient and optimal dialogical argument models finds its origin embedded in this intersection of the three technological domains. Dialogical argumentation has shown to to improve critical thinking and reasoning capabilities of the participants (Visser and Lawrence, 2022; Felton et al., 2015) in a discussion allowing for a richer and productive exchange. They can improve the learning and reasoning capabilities of a students, especially in a collaborative ecosystem. It is also widely accepted that argumentation enhances deliberative interactions to generate justifiable and reasonable consensus or a conclusion that is acceptable by all the parties (Schneider, 2014). This can be useful for making better artificial agents and find application in customer care services, content moderation, and decision making. Likewise, public policies and proposals that are built on informed participation (participants are aware of the nature of discussion and all of the arguments being presented to make an informed choice) can greatly improve the process of deliberative democracy (Lustick and Miodownik, 2000; Iandoli et al., 2018; Atkinson et al., 2006; Zenker et al., 2023).

Dialogical argumentation deals with the task of studying how arguments are structured, connected and presented in a dialogue. we begin by defining what constitutes as a dialogue, for which we refer to the definition proposed by Douglas Walton (Gordon and Walton, 2009; McBurney et al., 2010). A dialogue can be defined as an ordered 3-tuple $\langle O, A, C \rangle$ where O is the opening stage, A is the argumentation stage, and C is the closing stage. Each of these stages is marked by having distinct moves that can be made by the participants in the discussion. These moves are also restricted by the type of the dialogue being analyzed and thus are subjective from one context to another. For instance, there is no need to identify the position in an information-seeking dialogue (refer to Table 2). Likewise, deliberative dialogues can have a "brainstorming" event during their opening stage where various proposals are put forward in an attempt to underscore the available choices or courses of action. Efforts are being made to build a comprehensive list that defines these moves, based on speech acts, argumentative strategies (observed from debates and legal proceedings), and the nature of the discussion (Felton et al., 2022). In Table 1, we list some common linguistic and dialectical moves that participants can make during these stages, compiled from various works of Felton et al. (2022); Asterhan and Schwarz (2009); Hitchcock et al. (2001); Kok et al. (2011).

Stage of Dialogue	Move	Micro-Purpose	Sample statements
Opening	Initiate	start discussion	"I have a doubt..."; "I believe that...."
Opening	Identify	define issue and/or proposal	"We should..."; "We must... "; "We need to..."
Argumentative	Argument / Counter - A	advancing arguments	"I propose that..." ; "This promotes/destroys..."
Argumentative	Justify / clarify	invite elaboration of argument	"Please explain..."; "... needs to be clarified"
Argumentative	Add / advance	build / support argument	in addition.."; "...supported by..."
Argumentative	Counter / rebut	critically evaluate argument	"I disagree..."; "I challenge..."
Argumentative	Withdraw / concede	withdraw one's argument	"My argument is flawed..."; "I withdraw ..."
Closing	Recap	review / summarize	"To summarize / conclude ..."
Closing	Accept / reject	reaching consensus	"I end discussion...", "I accept / reject proposal ..."

Table 1: Common moves possible at each stage of a dialogue.

In the opening stage, participants agree to initiate or participate in a discussion on a particular issue/proposal/proposition. All participants identify the goal of the discussion (referred to as the collective goal) as well as their own individual goals that they want to advocate or achieve through the course of the discussion. This initial exchange and the goals themselves establishes the type

of dialogue that will take place in this interaction. In the Table 2, we list the seven commonly identifiable type of dialogues as proposed by Walton (McBurney et al., 2010; Parsons, 2007). Identifying the type of the dialogue, and the individual and collective goals, are critical for determining the type of discussion that will take place and accordingly the type of argumentative analysis that is required to understand the exchange. However, it should be noted that it is very natural for the nature of the dialogue to shift or transition from one type to another during the dialogical interaction as new information is introduced and exchanged between the participants. This dialectical shift generally occurs during the argumentative stage (Andone, 2008; McBurney et al., 2010). For instance, in an ongoing debate in a legislative assembly, there is a motion to pass on whether a dam should be constructed or not. Consequently they consult experts in engineering and ecology for advise. Thus, the dialogue naturally transitions from being a deliberative dialogue to an information-seeking one and then back to being deliberative as the available choices are again made clear and a consensus may be reached. This smooth transition or shift is referred to as dialectical embedding (Andone, 2008; McBurney et al., 2010) and establishes a productive and functional relationship between the two dialogues. These are often observed on formal discussions with the participants and goals being explicitly defined. Alternatively, these shifts can be abrupt, with a clear line of interference observable in the argumentation process. This can often be seen in informal discussions such as the ones on social media platforms where there is a variable level of participation and diversity of position that can seek to upset the flow of discussion.

Type of Dialogue	Initial Situation	Participant's Goals	Goal of Dialogue
Persuasion	Conflict of Opinions	Persuade Other Party	Resolve or clarify Issue
Inquiry	Need to Have Proof	Find and Verify Evidence	Prove (Disprove) Hypothesis
Discovery	Need to Find an Explanation of Facts	Find and Defend a Suitable Hypothesis	Choose Best Hypothesis for Testing
Negotiation	Conflict of Interests	Get What You Most Want	Reasonable settlement both can live with
Information-Seeking	Need Information	Acquire or Give Information	Exchange Information
Deliberation	Dilemma or Practical Choice	Co-ordinate Goals and Actions	Decide Best Available Course of Action
Eristic	Personal Conflict	Verbally Hit Out at Opponent	Reveal Deeper Basis of Conflict

Table 2: Seven commonly identifiable Types of Dialogue. Source: (Gordon and Walton, 2009)

Usually, dependent on the type of the dialogue, we can also observe the establishment of the global 'burden of proof' (Gordon and Walton, 2009) in the opening stage. According to Walton, burden of proof can be defined as, 'allocation made in reasoned dialogue which sets a strength (weight) of argument required by one side to reasonably persuade the other side'. There is a growing literature of the role and significance of studying burden of proof in argumentation and on formal dialogue models (Gordon and Walton, 2009; Godden and Wells, 2022). In simple words, the participant/party in the dialogue that fails to provide or resolve their respective burden of proof in the dialogue tends to be on the loosing or compromised side. Identifying and modeling this burden of proof is a critical area of research and focus in the argumentation society especially in the fields of law and public policy making, where proof or evidence play a critical role in winning the argument. Many argumentative tactics and strategies have been designed, centered around the burden of proof, in order to win the argument (Gordon et al., 2007; Godden and Wells, 2022).

In the next stage referred to as the argumentative stage, the participants or parties (participants sharing the same goal) establish their position on the issue or proposal. A commitment pool can be identified that contains each party's respective stance on the issue and the arguments that they hold critical to their defense. They take turns to make moves, identifiable by speech acts such

as asking questions, making assertions, or putting forward an argument, as listed in the table 1. As each party presents their argument, there is also an emergence of local burden of proofs associated with each statement. Through the course of the argumentative stage, there is exchange of the arguments and their burden of proofs between the parties. The local burden of proof for each argument can change and move from one side to the other as new arguments or evidence are put forwarded and critically questioned. As each party makes their move, there is a consequential result of insertions or retractions from each party's proposal/opinion from their commitment pool as they propose, acknowledge, concede or withdraw their argument. This dynamic exchange is vital for moderating the flow of discussion as well as ensuring that the discussion is progressing toward the original goal of the discussion. Any deflection from the main flow of discussion can be see as attempts to distort or misguide the discussion (Chang and Danescu-Niculescu-Mizil, 2019).

Finally, at the closing stage, we study the outcome of the argumentative stage and determine which party has successfully met their global burden of proof, as per the requirements set for it in the opening stage. Usually at this stage, moves such as summarize or conclude are enacted that underscores if the goal of the dialogue has been achieved and determine the outcome of the discussion.

It is important to establish that arguments in themselves are of defeasible nature, i.e. they can be defeated or proven inadequate in a later stage, as new information and arguments are provided. This defeasible nature of argumentation leads to two main types of dialogues of interest, persuasion and deliberation. These find immense application in the fields of political science, law, public policy administration, and scientific discourse. From the table 2, we can note the difference between the two in terms of the goals and the initial situation with which the discussion starts. Deliberative dialogues are not aimed at finding the truth but arriving at a decision on what should be done. They often start with identifying the problem or issue that has to be resolved, after which proposals are put forward by the parties (referred to as 'brainstorming') and they start establishing their respective positions or stance on the matter. These discussions are designed with collaboration as the central navigating force where the parties collectively steer the proposed actions toward a common goal, i.e. reach a consensus either through agreement or compromise. There are no explicit winners or losers in this form of discussion. Persuasive dialogues instead are truth-directed discussions where the proposition is either accepted or rejected by the parties. The discussion starts with a claim or proposition supported by one party and contested by another. The central navigating force in the discussion is the adversarial interactions that deal with proving or disproving the propositions until a resolution can be clearly made and the conflict is resolved. While both deliberative and persuasive dialogues differ in their opening and closing stages as well as in regard to their respective collective goals, it is interesting to observe that the argumentative stage in both are practically similar. The interactions observed during the argumentative stage is of great interest for the purpose of identifying the nature of support for a particular proposal or proposition as well as identifying the strong arguments that are placed by either side of the motion.

For the course of this chapter, we shall primarily focus on these two types of dialogical argumentation. In the next section, we briefly highlight some of the common schools of thought that identify the theoretical approaches to modeling these arguments. In the third section, we discuss

some current symbolic and representational approaches that have been built to study these dialogues, mentioning some software and tools that can be used to interpret and represent these dialogues. Finally we conclude the chapter by underscoring some of the challenges faced by the dialogical research community and the future directions that this research field is moving toward.

3.2 Modeling Dialogical Argumentation

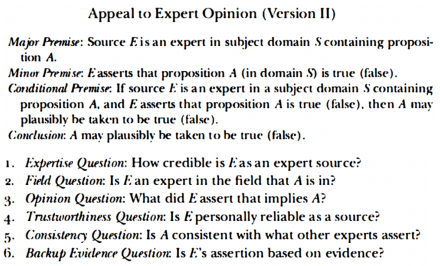
We start this section by stating that the theoretical approaches mentioned here are adapted from the theory of argumentation which in itself is vast and diverse in historical interpretations and theories. For the sake of brevity, we shall only focus on the aspects of these theories that are applicable to the dialogical domain. Dialogical argumentation was studied under the field of ”dialectics” that historically focused on studying argumentation in debates where two or more parties present diverse viewpoints or opinions and there is a desire to reach an agreement or consensus among the parties (Walton, 2007). It didn't receive much attention in the beginning as argumentation was primarily studied from the formal logic perspective. Dialectical argumentation was (and still rightfully is) put under the domain of informal logic, as opposed to well-studied formal logic (Walton, 2007) which studied argumentation from a logical and rhetorical viewpoint. A recent interest in studying informal fallacies (such as the ones noted in Aristotle's *Topics* and *On Sophistical Refutations*), shifted the focus on studying arguments away from their logical and abstract forms (as defined in argumentation schemes) and more toward conversational formats where the argument was no longer formal and abstract but rather contextual. This new school of thought required analysing the arguments by also taking into account related texts such as explanations and asking of questions, that otherwise had not been seen as argumentatively relevant. These statements played an essential role in the sequence of argumentation and in evaluating the strength of the argument, whether the premises support the conclusion as good reasons are not. We establish that dialogical argumentation falls under the informal logic and argumentation theory school of thought, referred to as dialectical reasoning. Given that informal logic dictates that the arguments are contextual in nature, we now present some novel schools of thought that have emerged which attempt to analyse and model these arguments from a practical reasoning viewpoint.

The ”pragma-dialectical” school proposes to set-up ”critical discussion” as the model for the argumentation, wherein the aim is for the discussants to resolve their difference in opinion (van Eemeren and van Haaften, 2023). They focus on the ”reasonableness” of an argument that is, avoid fallacies which tend to obstruct the goal of a critical discussion. This goal is based in the resolution of opinions, i.e. reach a consensus. However, this approach may also often view the goal to be to ”win” the argument rather than reach consensus. Rhetorical techniques, seen as ”strategic maneuvering” can be employed as long as they don't derail the discussion. However the issue in this approach is that it seems to insist that the discussants are both committed to winning the discussion as well as committed to reaching a consensus. In other words, it does not adequately differentiate between the different types of claims that people may argue for. A more in-depth and comprehensive focus on pragma-dialectical approaches can be found in these papers (Wen, 2016; Visser et al., 2020; van Eemeren and van Haaften, 2023).

Another quite popular school of thought, Walton's taxonomy of argument schemes (Walton

et al., 2008), focuses on defining the normative models for identifying and classifying arguments based on an abstract structure of the argumentation. These schemes represent common types of arguments used in everyday discourse. Walton's follows a form of "presumptive reasoning" synonymous with the concept of "defeasible" or abductive reasoning. As stated above, these arguments were categorized as fallacious originally in the logic textbooks. However these inherently defeasible arguments form a critical component of everyday practical reasoning. For example, arguments based on expert opinions are often used in social and intellectual institutions as expert testimonies, DNA evidence etc., representing a dominant form of evidence. This recent paradigm shift about rational argumentation has affected many fields of science and industry such as law, cognitive science, artificial intelligence, philosophy, biology, and any other areas where rational argumentation is centrally important. In Walton's book on Argumentation Schemes (Walton et al., 2008), many important types of argument schemes have been compiled and substantiated, highlighting their premises, conclusions and the set of critical questions that should be answered in order to recognize the argument's strength. They identified roughly 60 different argumentation schemes. Critical questions play a vital role in the definition of scheme, as well as in the development of argument modeling applications — computational or otherwise — with the purpose being to capture these critical questions in an appropriate way. To substantiate this with an example, let's discuss one of the relevant argument schemes, namely the 'Argument from Expert Opinion'. We present two main interpretations of the Argument to Expert Opinion in figures 3 and 4. The first figure displays the argumentation scheme with the set of critical questions, as proposed by Walton. The conditional premise represents the Toulmin warrant (Reisert et al., 2015) that helps gives the argument its backing. This argument has a defeasible *modus ponens* structure. In a given case, argument of this form could throw weight on the conclusion that the proposition A is plausible. However, if E is deemed to not be a credible expert, it would defeat this argument and undermining accepting A. In the second figure, a more explicit representation of this argument form has been presented where in the critical questions are explicitly built into the argument scheme, making it complete by itself. However, from a practical implementation perspective, the first representation of the argument scheme is more useful as strikes a nice balance of presenting the core requirements for the proposition to stand, while also allowing the users to select strategically between the critical questions, for probing the weak points of the argument. Also from the argument diagramming perspective, the first form is more attractive as it is easier to represent the general structure of the scheme and allow the analyst with the flexibility to include only the relevant critical questions, while the full set of critical questions can be retained in the meta information of the scheme and can be inferred later if required.

One would question how argument schemes can be normative if these practical arguments themselves are defeasible in nature and can be challenged by asking the critical questions. The proposed solution to this challenge is the concept of "profiles of a dialogue" (Krabbe, 2002). These represent a sequence of moves that represent only a small part of the long dialogue but are descriptive in identifying the patterns of the moves. This can thus be used to make assumption about the nature of the dialogue, whether it follows a normative scheme, or deviates from one, which can be diagnosed as faults, errors or fallacies. Another challenge to these schemes are the issues of enthymemes, that is implicit arguments (Walton and Reed, 2005). These can be mitigated by making attempts to justify the argument using the critical questions and the concept

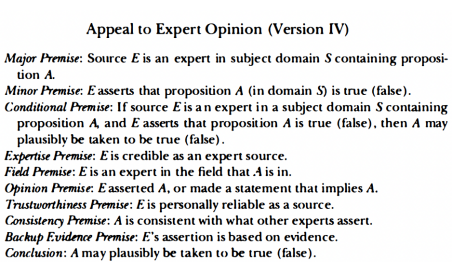


Appeal to Expert Opinion (Version II)

Major Premise: Source *E* is an expert in subject domain *S* containing proposition *A*.
Minor Premise: *E* asserts that proposition *A* (in domain *S*) is true (false).
Conditional Premise: If source *E* is an expert in a subject domain *S* containing proposition *A*, and *E* asserts that proposition *A* is true (false), then *A* may plausibly be taken to be true (false).
Conclusion: *A* may plausibly be taken to be true (false).

1. *Expertise Question*: How credible is *E* as an expert source?
2. *Field Question*: Is *E* an expert in the field that *A* is in?
3. *Opinion Question*: What did *E* assert that implies *A*?
4. *Trustworthiness Question*: Is *E* personally reliable as a source?
5. *Consistency Question*: Is *A* consistent with what other experts assert?
6. *Backup Evidence Question*: Is *E*'s assertion based on evidence?

Figure 3: Argument from Expert Opinion with separate critical questions (below). Source: Walton et al. (2008)



Appeal to Expert Opinion (Version IV)

Major Premise: Source *E* is an expert in subject domain *S* containing proposition *A*.
Minor Premise: *E* asserts that proposition *A* (in domain *S*) is true (false).
Conditional Premise: If source *E* is an expert in a subject domain *S* containing proposition *A*, and *E* asserts that proposition *A* is true (false), then *A* may plausibly be taken to be true (false).
Expertise Premise: *E* is credible as an expert source.
Field Premise: *E* is an expert in the field that *A* is in.
Opinion Premise: *E* asserted *A*, or made a statement that implies *A*.
Trustworthiness Premise: *E* is personally reliable as a source.
Consistency Premise: *A* is consistent with what other experts assert.
Backup Evidence Premise: *E*'s assertion is based on evidence.
Conclusion: *A* may plausibly be taken to be true (false).

Figure 4: Argument from Expert Opinion with critical questions as separate premises. Source: Walton et al. (2008)

of burden of proof. However, there is a need for rule that puts an end to the process of critical questioning and is often referred to as completeness problem for presumptive argumentation schemes. Another issue to be noted is that this model tends to fails in deliberative dialogues. In deliberation, proposals and their actions are studied as opposed to propositions. Following the presumptive approach, any action proposed by the participant that brings a benefit (that is, serves the goal) should be accepted, but in a scenario the participant has any more goals that counters this action, then that action should be rejected. Realistically, it is often the case that a participant will have multiple goals of differing nature, resulting in all of the arguments (actions) being defeated as per the presumptive reasoning. Thus, presumptive model of defeasible argumentation is not suitable for use and evaluation of deliberative arguments. Walton also highlights the challenges this approach can play in the field of artificial intelligence and computational modeling of arguments. As per him (Walton et al., 2008), argumentation schemes should be designed to be

- rich to cover a large proportion of natural argument
- simple so that it easily be taught, replicated and applied by students (at schools), as well as researchers for training AI models
- fine-grained so that it can be used both as a normative (generalized) and as an evaluative system
- rigorous and robust so that it can be adopted into various architectures and computational languages such as XML, JSON, etc.
- clear so that it is easy to be integrated into symbolic modeling approaches such as argument maps, argument graphs, etc.

Finally, we briefly mention another school of thought of argument modeling, namely Wageman's Periodic Table of Arguments. As highlighted in the works of Katzav and Reed (2004) and Hornikx (2013), most of the proposed argument schemes are unsuitable for use in the area of artificial intelligence due to the lack of formal ordering principle. For instance the presumptive approach to dialectic argumentation discussed above originates from an empirical starting point, rather than from a theoretical viewpoint, and thus is essentially not a complete set that even Walton them-self also acknowledge (Visser et al., 2021). At the same time the pragma-dialectical

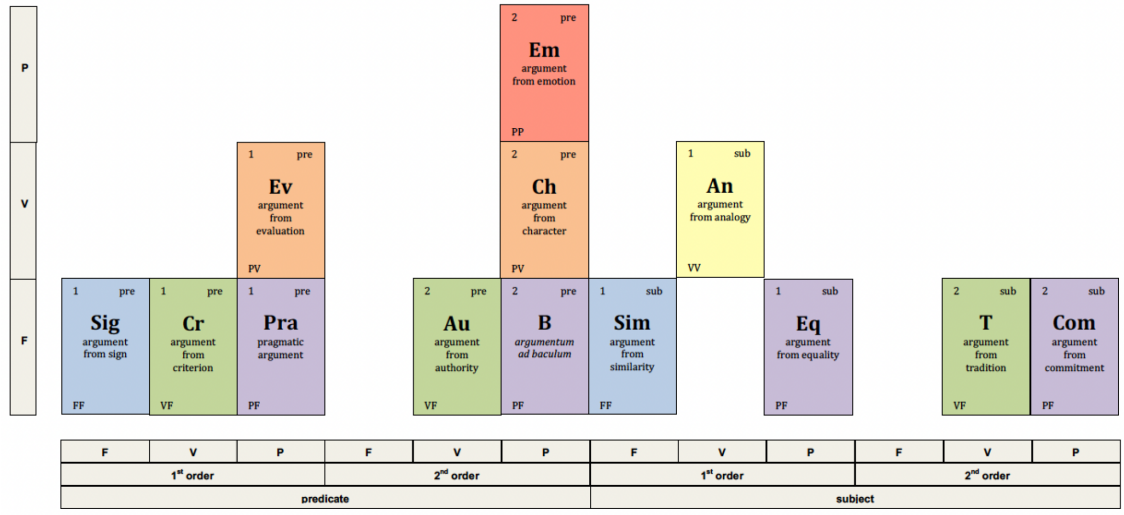


Figure 5: Periodic Table of Arguments. Source: Wagemans (2023)

approach, though being theoretical backed, doesn't seem to efficiently explain why there only three different types of argument schemes and not more or less. The three proposed schemes themselves fault in their interpretation (Hitchcock et al., 2001). Wageman proposed developing a classification of the arguments based on a set of formal ordering principles. The first principle establishes a distinction between the subject argument(s) and the predicate argument(s) based on a formal-linguistic analysis. Then, a distinction is made between first-order and second-order arguments based on the mechanism that governs argumentation from authority. Finally, the type of argument is characterised by the combination of the propositions they instantiate, the latter being based on the typology of propositions as proposed in the debate theory (Wagemans, 2014). These include the proposition of policy (P), the proposition of value (V), and the proposition of fact (F). Figure 5 presents the proposed Periodic Table of Arguments, based on these three principles. The author of the approach has indicated that this classification still needs to expand to systematically incorporate the dialectical and rhetorical accounts of the arguments. A newer version of this table can be seen in Visser et al. (2021).

3.3 Symbolic Approaches to Dialogical Argumentation

While the previous section focused on identifying and classifying arguments as a unit in a dialogue, this section focuses on studying relations between these units in form of symbolic approaches. We start by introducing some of the common theoretical approaches that lay ground to how arguments are connected to each other. The most prominent of which is the Inference Anchor Theory (IAT) (Budzynska and Reed, 2011) that provides a theoretical framework to dialogical argumentation. Built on the principles of Discourse Analysis and Argumentation Theory, IAT presents an explanation of the argumentative units in form of "anchoring" of the reasoning structures in persuasive dialogical interactions. It bridges the gap between logical reasoning and dialogical argumentation. IAT sets out to answer the question of where argumentation comes from in dialogical interaction, and acts as a theory-neutral scaffolding that integrates dif-

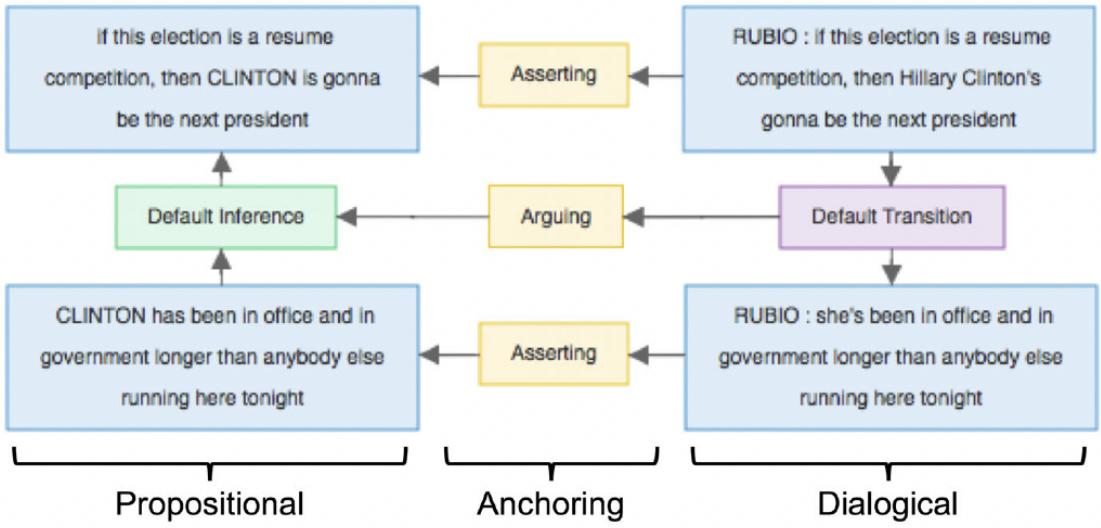


Figure 6: Diagrammatic visualization of a dialogical example from the US2016 corpus showing how propositional reasoning on the left is anchored to the dialogical realisation of the argument on the right. Source: Visser et al. (2021)

ferent communicative structures, namely dialogue structure, argument structure (including inference understood in the logical manner) and illocutionary forces such as asserting, suggesting or promising, to anchor argument structure in dialogue structure (Budzynska and Reed, 2011). IAT is also geared toward computational linguistics and software implementation. This is accomplished by adhering to the extended Argument Interchange Framework (AIF+) standard (Rahwan and Reed, 2009). The latter is a graph-based ontology that facilitates representations of different locutionary, illocutionary and propositional structures and allows for analysis of the argumentative discourse. AIF combines three different software aspects, namely argumentation protocols for example ASPIC+ with molecular arguments (Modgil and Prakken, 2014), software for visualization of arguments (like Rationale (Reid, 2011) or OVA (Janier et al., 2014)), descriptive logic matching tools for mathematical logic and IT-representation of knowledge (Rahwan et al., 2011). More information on how AIF can be used for modeling dialogical, more specifically deliberative, argumentation can be found in Rahwan and Reed (2009); Lisanyuk and Prokudin (2021). In figure 6, an example from the US2016 corpus (Visser et al., 2021) is given that employs IAT to study the relation between the different arguments proposed in the dialogical interaction. Other works of literature employing IAT and AIF include Lawrence et al. (2016); Hautli-Janisz et al. (2022).

Many software and tools have been developed that model dialogical arguments through various interactive visualizations, primarily as graphs or ontologies.

- OVA – developed by the Centre for Argument Technology of Dundee University (Scotland) (Janier et al., 2014);
- Carneades – developed by T. Gordon (Potsdam University) and D. Walton (Gordon et al., 2007);
- Rationale – initially developed by T. van Gelder's team in Melbourne University; today is a commercial software (https://www.rationaleonline.com/);

- bCisive – elaboration of Rationale for representation of argumentative support of decision-making (https://www.bcisiveonline.com/);
- VIANA - Visual Interactive Annotation of Argumentation developed by Fabian Sperrle and his team (Sperrle et al., 2019);
- BCause - developed by Lucas Anastasiou and his team for purpose of rich online discussions (Anastasiou and De Libbo, 2023);
- Kialo - developed by the team of Kialo for purpose of encourage critical discussions and debates (https://www.kialo.com/)

These tools have been used to automate the process of dialogical argument modeling, critical to automating the task. We see that these symbolic representations can be useful for visualizing argumentative dialogues and understanding how the dialogical process evolves over the course of the discussion. This allows for easier and faster annotation, relation identification, as well as determining the dialogical type and goals. Arguments can be connected to each other based on the the nature of their support, or based on their role in claim verification or conclusion generation, or based on their network features (such as the topic in discussion, the stance taken on the topic, as well as the speaker or proponent of the argument). These connections can be defined through various logical, dialectical, rhetorical, or discourse analysis techniques. This finds application in developing accurate and reliable resources for the purposes of computational argumentation as discussed in the next section.

3.4 Computational Approaches to Dialogical Argumentation

In this section, we briefly highlight some of the main dialogical applications and models that have been built based on the modeling of dialogical argumentation. Computational implementation of dialogical models fall under the domain of argument mining, where the objective is to retrieve or mine information/knowledge from the arguments present in the text (Budzynska and Reed, 2019; Arora et al., 2023). Mining from dialogical arguments can be grouped under three broad categories — argument structural analysis, argument content analysis, and argument network analysis.

In argument structural analysis, the objective is to identify arguments for the purpose of opinion mining (Rocha and Lopes Cardoso, 2017; Alhindi et al., 2020), fact-checking (Visser and Lawrence, 2022), and argument type identification (Wagemans, 2023). This automation often requires the task of defining an argument scheme as explained in the previous sections. The schemes is dependent on the objective of the task, like argument identification or argument classification, as well as on the source of the data, such as official public or scientific documents (Ruggeri et al., 2023), social media (Macagno, 2022), online debate and discussion forums (Abbott et al., 2016; Mestre et al., 2021; Goffredo et al., 2023), news platforms (Sardianos et al., 2015), or review submission platforms (Passon et al., 2018). From a dialogical context, there is an interest to study how different arguments are connected to each other as a relation to one other, such as a premise-claim relation pair (Boltuzic and Šnajder, 2016; **?**), or as a premise-conclusion relation pair (Gurcke et al., 2021).

In argument content analysis, the objective is to study the nature of the argument or arguments as a group. This finds application in tasks such as in argument strength estimation (Wachsmuth et al., 2017; Guo and Singh, 2023), stance (or sentiment) detection (Stefanov et al., 2020; Lai et al., 2020; ALDayel and Magdy, 2021; Alturayeif et al., 2023), opinion polarization (Belcastro et al., 2020; Kushwaha et al., 2022; Nguyen and Gokhale, 2022), and ideology analysis (Conover et al., 2011; Chen et al., 2017; Kawintiranon and Singh, 2022).

Lastly, in argument network analysis, there is an interest in developing expert models such as knowledge graphs (Chen et al., 2017; Pan et al., 2024) that are able to represent arguments as a network or a graph. This allows for improved knowledge representation and finds application in argument generation (Lawrence and Reed, 2017; Hinton and Wagemans, 2023), dialogical summarization (Misra et al., 2015; Egan et al., 2016; Misra et al., 2017), and argument search (Pan et al., 2024).

3.5 Conclusion

The first section introduced dialogical argumentation by defining what constitutes as a dialogue, marking the different moves that can be made at various stages of a dialogue, the different types of dialogue that can be defined in the discourse based on their initial situation, dialogic goals, and the individual goals. Another important concept of ”burden of proof” was introduced and discussed as an important defining element to the flow of discussion and estimating the conclusion of a discussion. A short comparison was made on two main types of dialogues, namely the persuasive and deliberative dialogues highlighting their differences and relevance in the discourse. From there, we proceed to the next section that focused on identifying and classifying the argument itself in the dialogic discourse. We discussed three main schools of thought, 'pragma-dialectical', 'Walton's taxonomy of argument schemes', and 'Wageman's Periodic Table of Arguments' that attempt to model arguments. For each of these schools, we presented some examples and applications in the literature as well as raised some identified concerns and challenges to the modeling approach. In the next section, we transitioned into the domain of symbolic representations in forms of graphs, argument frameworks and maps. This section focused on how different arguments in a dialogical interaction can be connected to each other and the relations between them was studied. This section also provided some existing software and tools that have been designed and released for the purpose of modeling dialogical arguments.

We started this discussion by mentioning three main technological advancements that have spearheaded the development of dialogic discourse, from an argumentative perspective. We conclude by stating some of the challenges and future directions observed in dialogical argumentation with respect to each of these fields.

Argumentation Theory. In section 2 and 3, we highlighted some of the underlying challenges when it come to defining dialogues from an argumentative perspective. Since dialogical arguments were originally seen as defeasible and belonging to informal logic, not much attention was given to them. This has resulted in lack of argumentation models or tools that are applicable to this domain. However, there is a recent change in the trend now, where focus is being navigated to this domain. Some prominent research events that actively make dialogical analysis as part of their research focus include conferences and special interest groups such as SIGDIAL

(https://www.sigdial.org/), COMMA (https://comma.csc.liv.ac.uk/), ECA (https://ecargument.org/) as well as workshops like the ArgMining workshop (https://argmining-org.github.io/2024/) and the ARGMAS workshop (http://www.mit.edu/~irahwan/argmas/). As we highlighted in section 3, there is also a growing trend in providing open-source and open-access resources such as dialogic data, software for modeling and visualization, as well as argumentation frameworks that can be used as template for modeling and building dialogic models. There is however, a need to ensure that there is consistency, transparency and reliability in how these resources interpret and represent the arguments (Visser et al., 2018; Lisanyuk and Prokudin, 2021).

Social Media. While in Deliverable 2 from the HYBRIDS Work Package 2: Discourse Analysis, we highlighted in brief how opinions and arguments in extension play a role in social media, we briefly reiterate the same. Argumentation models can find application in the social media platforms as moderation tools for detecting hate speech (Wilson and Land, 2021; Vecchi et al., 2021; Ladd and Goodwin, 2022), fake news (Kantartopoulos et al., 2020; Kotonya and Toni, 2019) as well as fact-checking (Alhindi et al., 2020; Visser and Lawrence, 2022; Hardalov et al., 2022). Likewise, supportive dialogical models can improve the quality of participation (informed participation) that is observed on these platforms when concerned with political discussion, electoral information exchange and campaigning, public policy administration. There is a need to develop fair and transparent models that can ensure that opinions are fairly represented on the platforms as well as in the models itself (Ghafouri et al., 2023; Rozado, 2023; Kobbe et al., 2020).

Conversational AI (agents). As conversational AI gains dominance as the lead source of knowledge generation, acquisition and exchange, it is imperative to build efficient dialogical models that can improve human-machine communication and interaction. A lot of new avenues are emerging that focus on how argumentation models can be integrated into large language models to improve their learning, reasoning and decision-making capabilities (Guo and Singh, 2023; Pan et al., 2024). There is immense scope for building expert models that can act as knowledge banks for determining which are best arguments to be made in a given context (Toledo-Ronen et al., 2016; Mou et al., 2024).

4 From Words to Functions: Analyzing Discourse through "Speech Acts Theory"

4.1 Speech Acts Theory from a Philosophy of Language Perspective

Historically, the positivist philosophy of language predominantly construed language as a mechanism for articulating factual assertions, perceiving the role of a statement as limited to describing a state of affairs, capable only of being true or false (Casalegno et al., 2003).

In the field of discourse analysis, where language is understood not merely as a system of communication, but as an evidence of aspects of society and social life (Taylor, 2013), the positivist approach is markedly inadequate. Viewing language as a mere tool for making factual statements, fails to accommodate the complex, interpretative processes that reveal how language constructs and is constructed by social realities.

In this regard, Wittgenstein (2019) introduced the notion that the essence of language lies not in its meaning but in its use, suggesting that language serves as a tool for social interaction. According to Wittgenstein, the pragmatic functions of language serves to achieve specific objectives within particular contexts through rule-governed "language games".

This perspective gave rise to Speech Acts Theory (SAT henceforth), which finds its roots in Austin (1975). In his seminal book *"How to Do Things with Words"* Austin distinguishes three level of speech acts analysis (for an overview see also to Harris and McKinney (2021); Nordenstam (1966)):

- **Locutionary**: refers to the actual utterance, i.e. the verbal, syntactic, and semantic components of any meaningful utterance. This level of analysis aims at answering questions like "what is being said?" (in terms of content), or "which communicative means have been used?" (in terms of tone of voice, grammatical structure). Analyzing the utterance "*You can't park your car there*" on a locutionary level could imply, for instance, detecting the sounds emitted to pronounce it (phonetic level), the syntactic elements of the sentence (phatic level, e.g. the subject-you, the main verb-park, etc.).

- **Illocutionary**: this level of analysis was introduced since any utterance (i.e. locutionary act) can be uttered with different *objectives*. This level of analysis, thus, aims at answering the question "in what way is this utterance being used on this occasion?". In this sense, someone could utters "*You can't park your car there*" to describe a local law or issuing a command, or even making a joke around, speaking sarcastically. It follows that performing an illocutionary act rather than other, in turn, will strongly depend on speakers intentions and the context in which the interaction is taking place.

- **Perlocutionary**: Austin further delineated illocutionary acts from perlocutionary acts, the latter being actions executed through the performance of illocutionary acts and contingent upon the subsequent effects of these acts. This level of analysis answers to the question "Which consequential effects upon the feelings, thoughts or actions of the audience is the speaker trying to obtain?". Referring to the car example, by commanding someone not to park there, one might perform this act to annoy him/her or to make them move their car.

Starting from this distinction, it emerges how in SAT, illocutionary acts occupy a central role due to the extensive range of functions and modes in which speech is employed, significantly influencing the nature of the communicative act. Therefore, it can be posited that the principal objective of SAT is to identify and delineate the illocutionary effects of a particular sentence within a given context.

One of the most widespread taxonomy of illocutionary acts is the one presented by Searle (1969). He organized the different illocutionary acts in five categories:

- Representative: the speaker aims at expressesing belief about the truth of a proposition.
- Directive: the speaker tries to commit the audience to perform an action.
- Commissive: the speaker commits himself to do sometng in the future.
- Expressive: the speaker express him/herself emotions and feeling (reactions) with respect a certain entity.
- Declarative: statements that directly alter the state of affairs in the world through their utterance. When delivered by someone with the appropriate authority or within the correct context, declaratives effectively change reality to align with the stated declaration.

In Fig.7 we provide an "at-glance" representation of the different levels at which Speech Acts can be analyzed. However, while the "macro-levels" —locutionary, illocutionary, and perlocutionary acts — serve as the foundational categories, it is essential to acknowledge that their "sub-level of analysis" can vary depending on the research context and objectives. This variability is symbolized by the inclusion of "*Others...*" under each category.

To date, in fact, several taxonomies have been developed to categorize the different illocutionary acts. This proliferation can be traced back to theoretical divergences or the need to specify the speech analysis according to the objective of the investigation.

With respect to the theoretical debates, Harris et al. (2018) provides an overview of the most relevant schools, which, depending on the case, ground the illocutionary acts' distinction on social conventions (Austin, 1975), intentions (Grice, 1969) and normativity (Sellars, 1969, 1954).

"Task-related taxonomies", on the other hand, are far more varied and numerous, precisely because of the manifold contexts in which SAT has been applied. This is especially apparent in recent uses of SAT to improve text characterization in Natural Language Processing (NLP) and Machine Learning (ML) for automated discourse analysis. These tools not only demonstrated the potential to expand SAT to new scalable operations, but also require to integrate SAT's theoretical constructs with the technical and methodological constraints of computational methods.

Consequently, this section aims to provide a comprehensive overview of NLP and ML most significant applications of SAT's. It will highlight how SAT's theoretical constructs have been operationalized in computational contexts, illustrating the dynamic interplay between theory and technology. This discussion will include specific examples and case studies that demonstrate the practical impacts of these theoretical applications, thereby offering insights into the evolving landscape of computational discourse analysis based on Speech Acts Theory.

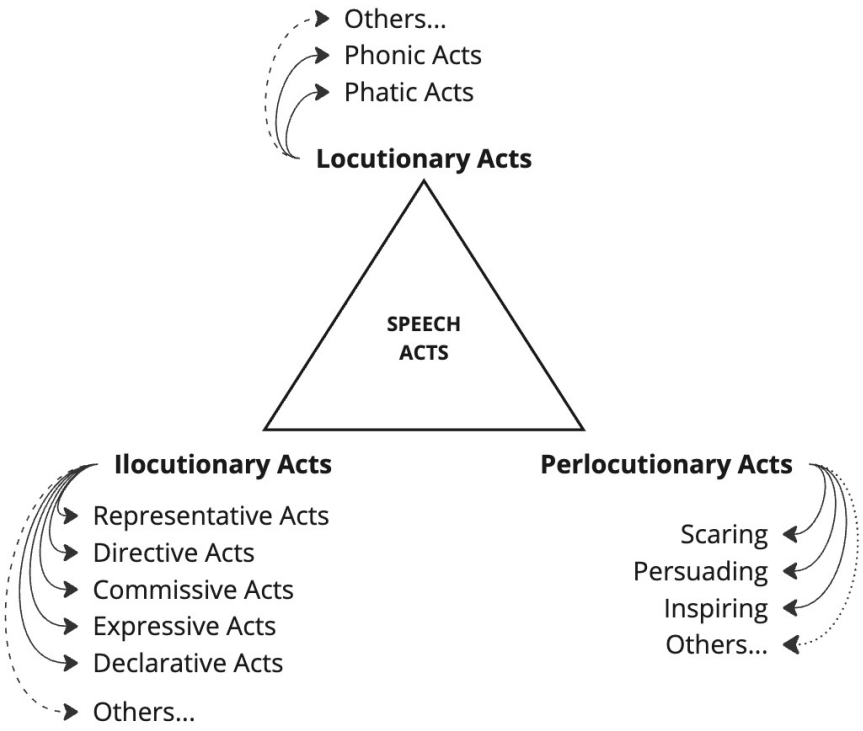


Figure 7: Tripartite Structure of Speech Acts, based on Austin (1975); Searle (1969)

4.2 NLP models for Automatic Speech Acts Classification

The objective of this section is to systematically organize and discuss the various NLP applications of SAT within the field of discourse analysis. With ”Automatic Speech Acts Classification” we refer to the multi-label classification task using SAT to to automatically characterize text excerpts based on their functional role within the linguistic interaction being analyzed.

The term ”classification” is employed because the labeling of a text excerpt is significantly influenced by the research objective. As briefly outlined previously, diverse taxonomies can be constructed according to the researcher's specific requirements. In this context, a ”Speech Act” does not possess a statute of ontological reality, allowing for it to be merely ”found” or ”detected”; instead, it is categorized based on the adoption of particular objectives and theoretical assumptions. This approach underscores the interpretative and pragmatic nature of speech act classification within the framework of discourse analysis.

To offer the clearest possible insight into SAT-related NLP applications, the papers reviewed in this section are categorized by the specific contexts of the analyzed discourses or the distinct objectives targeted.

4.2.1 Offline Interactions Analysis

A possible application of NLP methods for automatize SAT-based discourse analysis, is related to the computerized labeling of transcriptions of human interactions in real-life situations (Koo et al.,

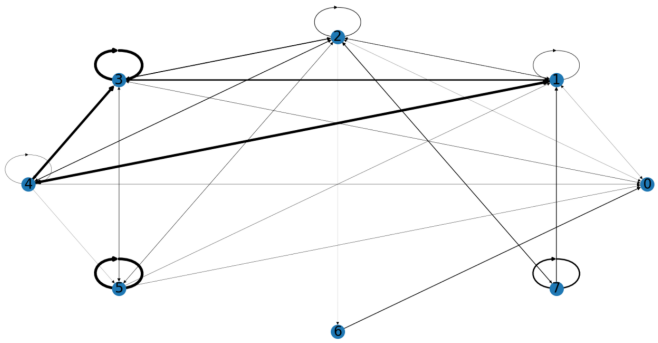


Figure 8: Classroom interactive sequence evolution (Schmidt et al., 2023).

2019; Kim et al., 2004; Ushio et al., 2017; Kim and Allan, 2019; Barbedette and Eshkol-Taravella, 2020).

Xia et al. (2022), for instance, focused on examining interactions between students and teachers within classroom settings. To facilitate this analysis, the researchers devised an enhanced Classroom Interaction Speech Act (CISA) coding system. This system refined the existing taxonomy to better suit the specific context of classroom discourse. Notably, it adapted the range of annotatable speech acts to reflect the distinct roles of speakers—incorporating, for instance, ”Lecturing” as a teacher-specific speech act, and ”Passive Response” and ”Noise” for student contributions.

On the computational side, to address the inadequacies of existing models that typically analyzed classroom dialogue on a sentence-by-sentence basis, the authors developed the Bert-TextConcat model. This model is a variant of the BERT model adapted to handle multiple texts or segments of classroom dialogue simultaneously.

Additionally, the data were processed by introducing special tokens to keep track of transitions between speakers (e.g., teacher vs. student) and different parts of the dialogue. These tokens help the model maintain context over the course of the conversation. This way the model was able to understand the flow of the conversation and the relational dynamics between speech acts, crucial for distinguishing different types of speech acts that might appear similar when taken out of context.

The performance metrics proved the hypothesis of the researchers: Acc[1]=81.24%. Moreover, thanks to this attention to the different turn-takings, the authors managed to generate a graphical representation of the interactions between different speech acts. In Fig.8, each node represent a specific speech act, while the edges (shown as directed paths) are used to describe their relationship. The thickness of the line indicates the number of interactions between two nodes; the direction of the arrows represents the order in which the different interactions occur.

4.2.2 Online Interaction Analysis

One of the most promising NLP application of SAT-based discourse analysis is the one related to the analysis of linguistic interactions taking place in different online environments. Digital plat-

[1] Accuracy (Acc) is a measure of how often the model correctly predicts the outcome, calculated as the ratio of correct predictions to the total number of predictions.

forms, as hubs for human online interaction, provide extensive datasets ripe for discourse analysis, particularly through the lens of SAT Moldovan et al. (2011); Cui et al. (2017); Zhang et al. (2011).

This subsection presents the work of Jin et al. (2022), focused on analyzing bragging practices in social networks (SNs). Among online interactions, in fact, SAT is particularly fitting to analyze the ones occuring in SNs given the tendency of these platforms to promote users to interact with others to craft idealized self-image (Michikyan et al., 2015; Halpern et al., 2017).

Bragging is defined as a speech act that either explicitly or implicitly attributes credit to the speaker for a positively valued attribute, such as a possession or skill cite (Rüdiger and Dayter, 2020). Despite constituting a pivotal act in politeness theory, bragging has been studied mainly through manual analysis of small datasets cite (Leech, 2016; Matley, 2018).

To automatize this analysis, Jin et al. (2022) devised two primary classification tasks to analyze bragging: (i) binary classification to determine if a tweet contains bragging or not, and (ii) multi-class classification to identify one of six bragging types (Achievement, Action, Feeling, Trait, Possession, and Affiliation) or non-bragging statements. The study utilized a dataset of 6,696 tweets, annotated through discussions until reaching a Krippendorf's Alpha above 0.80 (Jin et al., 2022).

The authors trained different models for both the tasks. In binary classification, transformer models significantly outperformed the baselines, with BERTweet excelling due to its training on English tweets. The addition of LIWC features enhanced performance further, indicating their efficacy in capturing linguistic elements characteristic of bragging (F1=72.42).

In multiclass classification, the BERTweet-Clusters model was particularly effective (F1=35.95), indicating that understanding the topical context of tweets enhances the identification of bragging types Jin et al. (2022).

To further elucidate which features most effectively predict bragging, the authors conducted a detailed linguistic analysis. They identified that certain unigrams, particularly personal pronouns and positive terms, significantly contributed to the model's performance. Additionally, the LIWC category "Achieve", which includes terms associated with accomplishments and success, was also predictive of bragging content. This last analysis highlights the nuanced understanding that computational methods bring to interpreting the pragmatics of language use in online settings.

4.2.3 Abusive Language

Phenomena such as insults, bullying, and discrimination have proliferated on social networks, prompting the development of automatic detection tools. However, such task present a complexity inherent to the use of colloquialisms and idioms that may not intend to offend, making it essential to move beyond basic lexical analysis. To address this issue, SAT could be used to define the necessary criteria for determining the presence of insulting, offensive, or aggressive language.

In this section, we discuss the work of Diaz et al. (2022), which utilizes Austin (1975) concept of illocutionary acts to develop an algorithm for automatically detecting offensive, vulgar, and aggressive language. Recognizing that the illocutionary force of a speech act is context-dependent (Fromkin et al., 2017), and considering the limited contextual cues within tweets, the researchers required annotators to use their sociopragmatic understanding to determine the illocutionary force

conveyed.

Consequently, they established pragmatic definitions for each studied insulting speech act - namely ”offensive”, ”aggressive”, and ”vulgar”. These definitions informed a flowchart of criteria that guided the annotation process, posing strategic questions to direct the annotator's decision-making. The flowchart facilitated the creation of a dataset comprising Spanish tweets categorized into the different classes of offensive language. The agreement score achieved a Kappa score of 0.91, which indicates strong annotation consistency due to the proposed schema.

Subsequently, a SVM model was trained to automatically detect these speech acts, achieving substantial performances: Acc=0.77 and F1=58. The authors emphasized that these results should also be interpreted in light of the inherent challenges SVM models face in comprehending the contextual nuances of sentences.

In this regard, we acknowledge also the work of Komalova et al. (2022), which, starting from the same issue of Diaz et al. (2022), elaborated linguistic criteria useful for the detection of ”insulting speech” using the definition provided by the Russian legislation. The authors crafted these criteria into vocabularies suitable for the training of computational models aimed at the automatic detection of such speech acts. Their empirical work utilized data from posts on the Russian social network VK.

In this instance, the integration of deep learning techniques with vocabularies grounded in human knowledge enhanced the performance of the models. This hybrid approach (Panchendrarajan and Zubiaga, 2024), in fact, resulted in a F1[2]=71, demonstrating the efficacy of combining advanced machine learning algorithms with contextually informed linguistic tools in the detection of insulting speech.

4.2.4 Political Analysis

The analysis of speech acts use in political communication has received only little attention despite their pervasiveness and utility in this communication context (Hashim and Safwat, 2015; Ulum et al., 2018).

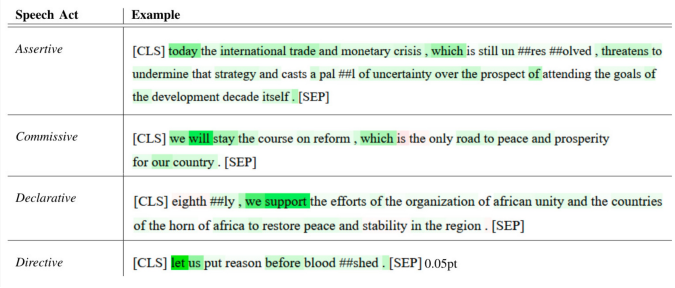
The network of deontic[3] moral forces within society, in fact, can be seen as established and maintained via particular types of speech acts. In this regard, Searle (1976) postulates that *declarative speech acts*, in particular, are crucial in forming institutions and institutional facts.

In light of this, the emergence of freely accessible and machine-readable digital resources constitutes an opportunity to use SAT to analyze extensive datasets of political speeches, such as parliamentary debates (Baturo et al., 2017) and addresses by individual politicians (Peters and Woolley, 2019).

To exemplify SAT possible application in the political domain, we present Schmidt et al. (2023). The authors created a dataset labeling the US State of the Union corpus and the UN General Debate corpus (UNGD) using Searle (1969)'s taxonomy. The labels included ”Assertive”, ”Expressive”, ”Commissive”, ”Directive”, and ”Declarative”, with the addition of a ”None” category for

[2]The F1 score is a measure of a model's accuracy, considering both the precision (how many selected items are correct) and recall (how many correct items are selected). It is useful for evaluating models, especially when the data has imbalanced classes.

[3]”Relating to moral ideas such as responsibility, permission, and obligation” (Cambridge Dictionary, 2024). Within philosophical discourse, this concept is often used to describe rules or laws that dictate what is allowed or obligatory in a given social context.



Speech Act	Example
Assertive	[CLS] today the international trade and monetary crisis , which is still un ##res ##olved , threatens to undermine that strategy and casts a pal ##l of uncertainty over the prospect of attending the goals of the development decade itself . [SEP]
Commissive	[CLS] we will stay the course on reform , which is the only road to peace and prosperity for our country . [SEP]
Declarative	[CLS] eighth ##ly , we support the efforts of the organization of african unity and the countries of the horn of africa to restore peace and stability in the region . [SEP]
Directive	[CLS] let us put reason before blood ##shed . [SEP] 0.05pt

Figure 9: Visualization of key features: positively contributing features in green, negatively contributing ones in red. Feature importance is indicated by brightness. Adapted from Schmidt et al. (2023).

utterances that do not fit into the predefined categories.

While this taxonomy facilitates a structured analysis of speech acts, we highlight how the inclusion of the ”None” category raises theoretical concerns. In fact, each utterance in a communication context presumably carries an illocutionary force; thus, categorizing an utterance as ”None” suggests a potential oversight in capturing all meaningful illocutionary effects within the dataset.

On the machine learning side, Schmidt et al. (2023) used a combination of active learning and supervised machine learning techniques. Austin (1975), in fact, identifies performative verbs as key indicators of illocutionary acts, hence providing a foundational linguistic model for the classification process. The study leveraged these linguistic indicators to inform the labeling process in a weak supervision setting, subsequently refined through an active learning one[4].

The authors employed a DistilBERT to reduce the computational overhead with minimal performance drawbacks compared to the resource-intensive BERT model. This way they managed to accelerates the iterative training processes. To address the often-criticized ”black-box” nature of such deep learning models, an ablation study was conducted. The authors systematically removed features to evaluate their impact on the model’s performance. As shown in Fig.9 by doing so, the authors managed to understand which linguistic features contribute most significantly to the classification task, enhancing transparency into the decision-making process of the model.

Finally, to illustrate the possible practical application of SAT application in the political domain, the authors carried out a temporal analysis of speech acts in political speeches, analyzing their variation over time in response to significant political events. The case study focused on the Ukrainian political crisis and measured shifts in communication strategies and rhetoric, providing insights into the dynamics of political discourse and the strategic use of language by political actors, as shown in Fig.10.

4.2.5 News Analysis

An additional domain of application for SAT is the one of ”news analysis”. da Silva et al. (2024), for instance, applied SAT to enrich the semantic representation of news texts, demonstrating

[4]Active learning employs query strategies to select informative samples from an unlabeled data pool for labeling, thereby improving learning efficiency. It involves a cyclical process of training a model on a labeled data pool to identify key unlabeled samples, which are then validated by a domain expert and added to the labeled pool. This continues until a set performance criterion is met.

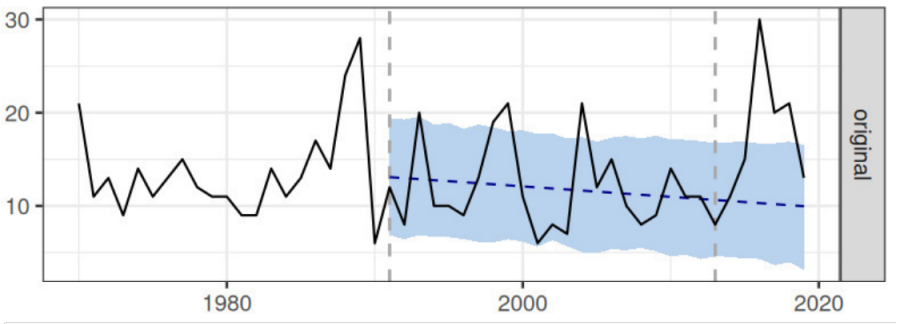


Figure 10: Speech Acts usage in absolute numbers between 1991 and 2019. Adapted from Schmidt et al. (2023).

how SAT can be used to identify the communicative intentions embedded in news discourse, facilitating a deeper understanding of the information conveyed.

One of the main points raised by the authors regards the challenges posed by the diverse taxonomies used in SAT classification. To overcome this issue, the authors emphasize the need to refer to international standards for labeling guidelines. They adopted the ISO 24617-2 standard (ISO, 2020), which differs significantly from the taxonomy derived from Searle (1969) (see previous section), including a more detailed division into 56 communicative functions across nine dimensions. This approach not only ensures compatibility with international standards, but allow also for a finer-grained analysis that accommodates the complexities of real-world discourse.

Using this schema, speech acts were manually labeled on a subset of the Porttinari-base corpus [5], a Brazilian Portugese dataset manually annotated with (morpho)syntactic features, under the Universal Dependencies (UD) paradigm (Nivre et al., 2020).

We emphasize how, in this annotation campaign, the authors explicitly dealt with a recent thorn issue of SAT: illocutionary pluralism. As stressed by Lewiński (2021), in fact, the same utterance token, in one unique speech situation, can intentionally and conventionally perform a plurality of illocutionary acts. To address this issue, da Silva et al. (2024) instructed the annotators to prioritize the assignment of the most *specific* communicative function available. While constituting a fist step towards a more meticulous annotation strategy, this solution strongly relies on annotators personal understanding both of the definition of the speech act and the contextual value of the analyzed text. This, in turn, underscores the inherent complexity of classifying speech acts, where the boundaries between categories are not always clear-cut.

Subsequently, the authors trained a machine learning model using BERTimbau, a variant of BERT adapted for Brazilian Portuguese. In doing so, the study also confronted issues inherent in the BERT architecture, particularly the challenge of dealing with class imbalance. Some speech acts, in fact, are significantly less frequent in the dataset. This imbalance skewed the model's ability to accurately classify less common speech acts, reflecting a common limitation in applying deep learning to real-world linguistic data. The authors approached this issue modifying the loss function to penalize misclassification of minority classes more heavily, trying to balance the influence of dominant classes in the training process.

Overall, the BERT architecture, fine-tuned to this specific task, highlights the adaptability and

[5] https://github.com/UniversalDependencies/UD_Portuguese-Porttinari

effectiveness of transformer-based models in handling the nuances of language specific to different contexts (Acc= 91.6; F1=91.4).

4.2.6 Health Promoting Communication

Another promising domain for SAT NLP application is the one of health promoting communication. Despite constituting an intriguing niche of computational linguistics, different authors addressed this topic Zhang et al. (2011); Laurenti et al. (2022); Epure et al. (2017); Cui et al. (2017). In this section we present the work of Laurenti et al. (2022), aimed at delving into the dynamics of how speech acts on social media can illuminate the urgency of communications during crises such as natural disasters, thus providing emergency responders with critical insights that could streamline their real-time responses.

The researchers have developed an innovative two-tiered taxonomy for categorizing speech acts in tweets. At the "*tweet level*", speech acts were classified holistically, capturing the overall intent of the entire message. The categories at this level include "*Assertive*", defined as Searle (1969); "*Subjectives*", which express personal feelings or evaluations; "Jussives", which can be assimilated to the "Directive" of Searle (1969); and "Interrogatives", which are questions posed by the tweeter. At the "sub-tweet level", the classification becomes more granular, dissecting segments of the tweet to identify finer nuances in communication. This level includes categories such as "*Open-Options*" and "*Other-Jussives*" under "*Jussives*", or distinctions between "*Reported*" and "*Proper Assertions*" under "*Assertives*".

To automatically detect these speech acts, the authors tested a mix of deep learning models. Among the others, they tested BERTbase for its robust multilingual capabilities, FlauBERTbase and CamemBERTbase for their French-specific contextual embeddings. Notably, the CamemBERT models were also experimented using focal loss to better handle class imbalances, returning the overall best performance (Prec.=75.66; Recall=71.95; F1=73.55)[6].

Interestingly, the research uncovered "*Assertive*" speech acts are predominantly linked with *urgent messages that necessitate immediate action*, thus highlighting the critical information within crisis communications. In contrast, "Subjective" speech acts often correlate with less urgent content, reflecting personal reflections that, while valuable, may not require immediate response. Finally, the roles of "*Jussive*" and "*Interrogative*" speech acts demonstrated variability in urgency, suggesting that the context of the crisis significantly influences the interpretation of these communications.

4.3 Conclusions

This section has provided a comprehensive examination of Speech Acts Theory. The exploration began with the philosophical underpinnings of SAT, highlighting its emergence as a response to the limitations of earlier language theories that did not account for the pragmatic aspects of language use. We discussed the three core types of speech acts—locutionary, illocutionary, and perlocutionary.

[6] For the description of these metrics see the footnote in Sec.4.2.3

Further, we detailed the application of SAT within NLP and ML, showcasing its utility in classifying speech acts across various communicative contexts. While our focus has centered on areas most pertinent to the social sciences, such as abusive language detection and political discourses analysis, it's important to note that we have not delved deeply into more operationally oriented applications, like mail classification (Cohen et al., 2004).

On the technical front, our survey allowed us to conclude on the superior capability of deep learning models in capturing and interpreting the complexities inherent in human linguistic interactions.

These models, despite their potential, are often criticized for their "black box" nature, which obscures their internal workings. However, we have also demonstrated how the theoretical frameworks of SAT can be used to enhance the explainability of these algorithms. In Sec.4.2.4, for instance, the work of Schmidt et al. (2023) and Komalova et al. (2022) exemplifies how, by harnessing the insights of SAT, it is possible to guide deep learning models not only to predict, but also to rationalize their classifications in ways that are intelligible and firmly rooted in human communication principles.

Similarly, in Section 4.2.2, Jin et al. (2022) demonstrates how using a vast amount of empirical data, gathered and analyzed through computational models, can illuminate the grammatical and semantic information utilized by the machine to make its predictions.

In conclusion, while this brief survey underscores the extensive potential applications of SAT through NLP methods, it also illuminates several critical challenges. Therefore, we conclude this section by enumerating these challenges and leveraging them to outline future research directions in this field.

4.3.1 Limitations and Future Directions

- *Diversity and Dispersion in Annotation Schemes:* we highlighted the rich diversity of contexts in which SAT has been applied. However, this diversity has also resulted in a diaspora of differing annotation practices, which complicates comparative studies and contributes to a fragmented research landscape.

  - As pointed out by da Silva et al. (2024), one potential resolution to the issue of this critical pluralism is the adoption of international standards, such as ISO 24617 ISO (2020); Bunt et al. (2020). This standard could serve as a common reference for researchers, providing consistency across various levels of analytical granularity.

- Additionally, we note that this diversity in annotation practices often arises from the need to adapt analyses to specific contexts. For instance Xia et al. (2022) developed new speech acts tailored to roles like teacher or student. In this regard, another feasible solution could involve consistently clarifying the broader category, typically rooted in the taxonomy of Searle (1969), from which the "new speech acts" are derived. Essentially, this approach would involve tracing each specific taxonomy back to its original, more general framework. In this regard we signal Koo et al. (2019), who provides a good example of how to hierarchically organize the context-specific taxonomy in relation to the general one.

- *Illocutionary Pluralism:* as underscored by Lewiński (2021), a single utterance may encompass multiple illocutionary forces, contingent upon the context. This multifaceted nature poses a significant challenge to contemporary computational models, which typically attribute a singular illocutionary role to each utterance.

  - Future research endeavors should concentrate on devising coding schemes equipped to address this complexity. Such schemes would necessitate more stringent guidelines and explicit criteria to facilitate more precise and consistent annotations by researchers.

  - Additionally, future developments could aim to enhance models to recognize and categorize multiple illocutionary forces within a single utterance, thereby refining the granularity and accuracy of speech act analysis.

- *Under-represented Speech Acts:* as stressed by da Silva et al. (2024), the predominance of certain speech acts in training datasets can bias ML algorithms, leading to under-performance in detecting less common acts.

  - Further research should aim to balance datasets and refine algorithms to better represent and recognize a broader spectrum of speech acts. In this regard, *synthetic data* could be generated to artificially generate under-represented speech acts, thus enriching training datasets.

  - On the ”learning side”, techniques like ”transfer learning” (pre-trained models fine-tuned on targeted datasets including under-represented speech acts) or ”active learning” (a strategy where the model identifies gaps in its learning and requests additional data on under-represented speech acts) can be used to enhance the effectiveness of models trained on few data (as shown in Schmidt et al. (2023)).

- *Hybrid Methods:* as previously highlighted, hybrid methods that integrate deep learning with knowledge-based approaches can mitigate the opaque nature of these models. While various studies discussed within this review explicitly aimed at enhancing algorithms' transparency, additional efforts are essential to further promote and expand the adoption of such methodologies.

  - Future research should develop and refine comprehensive ontologies that capture the full complexity of speech acts across different languages and cultural contexts. These ontologies would provide a structured framework that can be used to train more nuanced and globally applicable models. This initiative requires concerted efforts both computationally and theoretically; the latter is particularly crucial as it can provide valuable insights into the types of information that should be embedded within the algorithm to enhance its functionality and applicability.

5 Discourse Analysis Detection of Conspiracy Theories Using NLP Computational Techniques

5.1 Introduction to Conspiracy Theories in Discourse Analysis

- **Definition and Characteristics of Conspiratorial Discourse:** Conspiracy theories are a form of alternative knowledge that challenge mainstream accounts of events, often implicating sinister powerful groups in these alternate narratives (Brotherton, 2015). As highlighted by Uscinski (2018), these theories are defined by their complex narrative structures that feature clear protagonists (the conspirators) and antagonists (those threatened by the conspiracy), simplifying complex realities into a more digestible and emotionally engaging story. These narratives often gain traction by exploiting societal fears and uncertainties, providing a seemingly coherent framework to otherwise random or disconnected events.
- **What's the Conspiratorial Language? — Discussion on the Specific Linguistic Features and Rhetoric Commonly Used in Conspiracy Theories:** The language of conspiracy theories is marked by its use of emotionally charged rhetoric, absolutes, and dichotomous thinking, creating an 'us versus them' scenario (Miani et al., 2021). According to Douglas and Sutton (2008), this language is not only persuasive but also designed to evoke strong emotional responses from the audience, reinforcing group identity and solidarity among believers. Common features include the use of loaded terms, hyperbolic expressions, and an emphasis on secrecy, exoterism, and revelation, all of which are crafted to sow distrust against purported enemies and validate the conspiratorial viewpoint.

5.2 History of Combating Misinformation: Overview of computational techniques to detect general misinformation

This section reviews techniques for detecting online misinformation before the emergence of large language models (LLMs), according to (Chen and Shu, 2023b). Detection methods are categorized into seven classes based on real-world scenarios:

1. **Capturing Linguistic Features**: Techniques focus on stylistic, complexity, and psychological features to differentiate misinformation from true information (Antypas et al., 2021; Rubin et al., 2016). Misleading content often exhibits longer length, limited vocabulary, negative sentiment, informal language, and exaggerated expressions (Chen and Shu, 2023b).
2. **Leveraging Neural Models**: Neural models, including Long Short-Term Memory (LSTM), Convolutional Neural Network (CNN), and more advanced models like Bidirectional Encoder Representations from Transformers (BERT), are used for feature extraction and prediction (Kalchbrenner et al., 2014; Kaliyar et al., 2021). These models replace manual extraction of linguistic patterns and offer improved performance.
3. **Exploiting Social Context**: Incorporating social context, such as user interactions and social networks, enhances misinformation detection. Research has found that user-news

interactions differ between fake and authentic news (Shu et al., 2019). Consequently, several studies have explored using social engagements as valuable auxiliary information for detecting misinformation (Chowdhury et al., 2020; Del Tredici and Fernández, 2020; Li et al., 2020). Methods include analyzing social engagements and leveraging graph-based models to capture dissemination patterns on social media (Bian et al., 2020).

4. **Incorporating External Knowledge**: External knowledge sources like knowledge graphs and evidential texts aid in verifying the authenticity of information. Knowledge graphs contain a large number of entities and their relationships, which are useful for verifying the accuracy of articles (Ciampaglia et al., 2015; Cui et al., 2020). Evidential texts refer to factual content that can be used to verify the authenticity of articles (Akhtar et al., 2022; Chen et al., 2022).

5. **Enhancing Generalization Ability**: To address the evolving nature of misinformation, research focuses on improving detectors' generalization under domain (Huang et al., 2021; Liu et al., 2024a) and temporal shifts (Hu et al., 2023; Zhu et al., 2022). This is important because misinformation can vary significantly depending on the context and can change rapidly over time. Techniques to enhance generalization include reinforcement learning-based domain adaptation, which allows models to adapt to new domains by learning from different but related datasets. In other words, this technique helps models adapt to new domains by learning from data that is different but related to what they've seen before. Using reinforcement learning, models are trained to improve their performance by interacting with various datasets, gaining rewards for correct predictions, and improving over time. Additionally, methods that use forecasted temporal distribution patterns help models anticipate and adjust to future changes in the nature of misinformation. Basically, this method involves predicting future trends in misinformation. By analyzing historical data, models can identify patterns that suggest how misinformation might evolve, allowing them to adjust in anticipation of these changes.

6. **Minimizing Supervision Cost**: Due to the difficulty of obtaining supervision labels, approaches like data augmentation (He et al., 2021), active learning (Farinneya et al., 2021), prompt-based learning (Huang et al., 2023), and weak supervision leaning signals are explored (Yue et al., 2023). Data augmentation is a technique that generates additional training data by making various modifications to the existing data. He et al. (2021) explores how this can be used to enhance rumor detection. Active learning is a method where the model selectively queries a human annotator to label the most informative data points. Farinneya et al. (2021) studies active learning in the context of minimizing supervision costs. Prompt-based learning involves using pre-trained language models with specific prompts to generate predictions, which reduces the need for extensive labeled data. Huang et al. (2023) discusses meta-learning with prompts to improve efficiency. And weak supervision learning leverage noisy, limited, or imprecise sources of supervision to train models. This can include using heuristics, domain knowledge, or other indirect signals. Yue et al. (2023) explores meta-adaptive learning using weak supervision signals. Early misinformation detection is also a significant focus, as detecting misinformation early is crucial to mitigate its impact before it spreads widely (Li et al., 2022b).

7. **Fusing Multilingual and Multimodality**: Combining multilingual and multimodal information is increasingly important. Research aims to leverage high-resource languages for low-resource ones (Chu et al., 2021) and integrate various modalities (text, images, video) for comprehensive detection (Abdelnabi et al., 2022). The detection of conspiracy theories in German-language platforms like Telegram presents unique challenges due to linguistic and cultural nuances. As discussed by Pustet et al. (2024), unlike models trained with keyword-based English datasets, models developed for German texts must address token-level bias introduced by language-specific features. This is true also for other non-English models. They demonstrate the potential of both supervised fine-tuning and prompt-based approaches using models like BERT and GPT variants, showing promising results in detecting nuanced and implicit conspiracy narratives without relying on keyword filtering Pustet et al. (2024).

These methods collectively contribute to a robust framework for detecting and combating misinformation across different platforms and contexts.

5.3 Challenges in Analyzing Conspiracy Theories with NLP

- **Complexity of language and subtleties in conspiracy theories.** The complexity of conspiracy theories presents substantial challenges for NLP analysis, as noted by Shahsavari et al. (2020). The nuanced and speculative language used in these theories requires sophisticated data collection and annotation techniques to accurately capture the underlying meanings and implications within massive volumes of online data. The COCO dataset, as we will later see, highlights the challenges in data collection and annotation specific to conspiracy theories. Collecting data from social media platforms, particularly Twitter, involves sifting through vast amounts of noise to find relevant content. Annotating this data accurately is critical to ensure the effectiveness of NLP models Langguth et al. (2023).

- **How to Annotate It?** — Strategies and Guidelines for Annotating Conspiratorial Language to Facilitate Effective Analysis: Effective annotation of conspiratorial language requires a clear set of strategies and guidelines that account for the unique characteristics of this discourse. Annotators should be trained to recognize not just explicit statements but also the more implicit cues, such as thematic consistency with known conspiracy theories, the presence of non-mainstream sources, and the use of language that seeks to undermine official narratives. It is also crucial to develop a robust tagging system that can accommodate the complexities of these narratives, possibly incorporating labels for different types of rhetorical devices and logical fallacies commonly found in conspiracy theories. This detailed approach helps in building NLP models that are more accurate in detecting and analyzing conspiratorial content across diverse digital platforms.

  Furthermore, training automated systems to effectively discern and categorize conspiracy theories presents additional complexity. Several misinformation datasets, such as the ISOT dataset and the LIAR dataset, have been released Ahmed et al. (2017); Wang (2017). These datasets typically classify information on a binary true/false scale or on a continuum

Table 2 The number of times each label was assigned

Category	Unrelated	Related	Conspiracy	Agreement (%)
Suppressed cures	3410	15	70	98.11
Behavior control	3160	167	168	92.90
Anti vaccination	3095	191	209	92.27
Fake virus	3009	178	308	91.13
Intentional pandemic	2905	122	468	85.61
Harmful radiation	3370	63	62	97.94
Depopulation	3187	56	252	95.11
New world order	3189	43	263	94.39
Satanism	3412	35	48	97.45
Esoteric misinformation	3322	75	98	96.39
Other conspiracy theory	2133	413	949	75.85
Other misinformation	3220	60	215	90.01
Total	908	790	1797	92.27

Most tweets are unrelated to most categories. Note that Overall does not refer to the sum of labels, but to the total number of tweets per class. Agreement refers to the inter-annotator agreement, and total agreement is the average agreement

Figure 11: Table with the number of times each label was assigned Pustet et al. (2024).

from true to blatantly false. However, this simplistic categorization fails to capture the nuances between different misinformation narratives, which is crucial especially in contexts like the COVID-19 pandemic, or the Great Replacement Theory where diverse and often contradictory misinformation narratives proliferate on social networks.

To address the complexity of conspiracy theories within misinformation narratives, especially notable during the COVID-19 pandemic, Langguth et al. (2023) developed a specialized dataset. This dataset includes a comprehensive annotation of 3,495 tweets, categorized into 12 distinct conspiracy theory narratives. Each narrative represents a main thread or variation of the same overarching conspiracy theory. Tweets are classified into one of three classes within each of these 12 categories, resulting in a total of 41,940 labels. Each tweet was labeled by three annotators to ensure reliability. The final label for each tweet was determined by a majority vote among the three annotators. This detailed categorization enables the machine learning classifiers to distinguish between narratives that may be related or even directly contradictory, facilitating a deeper understanding of how such conspiracy theories proliferate and evolve on social platforms.

5.4 Computational Approaches Specific to Conspiracy Theories

Conspiracy theories have been around for a long time, but with social networks and messaging services, they now spread faster and more widely than ever before (Pustet et al., 2024). Acknowledging this rapid dissemination, various studies have focused on devising computational

models for disentangling conspiracy theories and detecting threat elements within them (Shahsavari et al., 2020). Among these, Rigoli (2022)'s Computational Model of Conspiracy Theories (CMCT) stands out for integrating computational psychology and Bayesian decision theory to understand the psychological factors influencing the endorsement of conspiracy theories Rigoli (2022). This approach provides a different perspective by examining how prior beliefs, novel evidence, and expected consequences influence individuals' acceptance of conspiracy theories.

Given that the CMCT does not utilize NLP, there remains a significant need for additional computational models that apply NLP techniques to detect and analyze the linguistic and discourse features of conspiracy theories effectively. Event relation graphs have also been integrated into conspiracy theory identification, with the development of an event-aware language model to augment basic detection methods Lei and Huang (2023). Additionally, efforts have been made to measure the diffusion of conspiracy theories in digital information spaces, particularly in relation to emerging COVID-19 conspiracy theories on social media and news platforms Heft and Buehling (2022). Building on these foundational models, further researches have explored the use of emotion analysis in conspiracy theory detection, with the development of ConspEmoLLM, a model that combines emotions and instruction-tuning to identify conspiracy theories Liu et al. (2024b). This approach seems very sound, with the emotional features helping the detection of such conspiratorial discourses. Plus, studies have focused on improving machine learning algorithms to better detect conspiracy theories on social media, using text-mining techniques to identify patterns in conspiracy theory language Marcellino et al. (2021). Overall, due to the huge amount of daily content posted online, computational methods play a crucial role in the detection and analysis of conspiracy theories, with ongoing efforts to enhance detection capabilities and understand the spread of conspiratorial beliefs in digital spaces.

Conspiracy theories, particularly in digital environments, are marked not just by their content but by their spread across languages and platforms, which complicate detection and analysis. Recent studies, like the one by Pustet et al. (2024), emphasize the need for computational tools capable of analyzing conspiracy discourse across diverse digital ecologies. Their approach using multilingual and cross-platform NLP models showcases how computational techniques are crucial in capturing the diffusion and prevalence of conspiratorial content in non-English languages on platforms like Telegram. In fact, as it usually happens, the research is already taking some good steps with the English language models, but there is still much research needed in non-English language models (Pustet et al., 2024). Haupt et al. (2023) and Platt et al. (2022) both demonstrate the potential of machine learning and natural language processing in detecting and differentiating conspiracy language, with Haupt et al. (2023)'s study emphasizing the importance of Hybrids intelligence techniques that combine these techniques with qualitative content coding, with Dascălu and Dascalu (2014) providing a broader overview of the role of computational discourse analysis in understanding cohesion and coherence in text, which is particularly relevant in the context of conspiracy theories. Chong et al. (2021) presents a real-time platform for contextualized conspiracy theory analysis, which could be a valuable tool for researchers and policymakers in monitoring and addressing the spread of conspiracy theories. Parallel to these developments, there has been an increasing focus on the potential of Large Language Models (LLMs) for misinformation detection. Initially, some works have investigated directly prompting models such as GPT-3 (Buchholz, 2023; Li et al., 2023), InstructGPT (Pan et al., 2023), ChatGPT-

3.5 (Bang et al., 2023; Caramancion, 2023), and GPT-4 for this purpose (Pelrine et al., 2023). For example, Pan et al. (2023) introduced a program-guided fact-checking framework that leverages the in-context learning ability of LLMs to generate reasoning programs for veracity verification.

Similarly, Chen and Shu (2023a) studied ChatGPT-3.5 and GPT-4 using both standard prompting ("No Chain of Thoughts") and zero-shot chain-of-thought ("CoT") prompting strategies for detecting human-written and LLM-generated misinformation. Their extensive experiments show that the "CoT" strategy generally outperforms the "No CoT" strategy. But, since the knowledge contained in large language models (LLMs) may not be up-to-date or sufficient for detecting factual errors, some works have explored augmenting LLMs with external knowledge or tools for misinformation detection (Cheung and Lam, 2023). Furthermore, Shahsavari et al. (2020) utilize advanced computational techniques to delve into the narrative structures that underpin conspiracy theories related to COVID-19. Employing automated machine-learning methods, their research involves crawling social media sites and news reports to discover underlying narrative frameworks that support the generation of these conspiracy theories. By systematically mapping how various narrative elements — such as actants (key players such as individuals, organizations, and locations) and their interactions — are articulated within these frameworks, they offer detailed insights into how such conspiracy theories propagate through social media. The study showcases the use of a narrative framework model, which identifies the actants and the relationships between them that are recurrent in the storytelling related to the pandemic. These frameworks help in recognizing how different conspiracy narratives align with broader media reporting on the pandemic. The computational process is capable of monitoring these alignments in near real-time, which is pivotal for understanding and potentially countering the rapid spread of misinformation.

Moreover, the study by Lei and Huang (2023) introduces an innovative method that utilizes an event relation graph to detect conspiracy theories in long news documents. This approach models the relationships between events in an article to identify typical conspiratorial patterns, such as the unnatural linking of unrelated events or distorted presentations of event relationships, which are common in conspiratorial narratives Lei and Huang (2023). Their methodology improves the precision and recall of conspiracy theory detection, illustrating the potential of graph-based NLP techniques in understanding complex misinformation narratives. Additionally, the study by Pustet et al. (2024) illustrates the use of Large Language Models (LLMs) like GPT-3.5 and GPT-4 in a zero-shot learning context to detect conspiracy theories effectively in German Telegram messages. These models, particularly GPT-4, exhibited strong performance, achieving F1 scores comparable to those of supervised models trained on more narrowly defined English datasets Pustet et al. (2024). This highlights the growing utility of LLMs in language-agnostic applications - the ability of a tool, model, or application to function effectively across different languages without being tailored specifically to any single language - and supports the need for further development in multi-lingual NLP tools.

5.5 Future Directions and Research Opportunities

As these studies demonstrate, the field is poised for further advancements that promise to refine our understanding and detection capabilities. For instance, the works of Pustet et al. (2024); Cheung and Lam (2023) suggest a promising direction for future research in improving the accu-

racy of conspiracy theory detection across diverse digital platforms and languages. Their studies underscores the potential for expanding NLP tools that can adapt to the varied and evolving nature of digital discourse, thereby enhancing the robustness of computational models against the dynamic backdrop of conspiracy theories online.

5.6 Conclusion

Summary of the Research Importance and Societal Implications

The analysis of conspiracy theories using NLP computational techniques is crucial for understanding the proliferation and impact of these narratives in digital spaces. Conspiratorial discourse, characterized by emotionally charged language and complex narrative structures, poses significant challenges to societal trust, institutional credibility and democratic processes. By leveraging NLP techniques, researchers can decode the linguistic and rhetorical patterns of conspiracy theories, providing insights into their spread and influence. This research highlights the need for sophisticated computational models to detect and counteract misinformation, ultimately contributing to a more informed and resilient society.

Contribution of Findings to Strategies Against Misinformation

In conclusion, by categorizing and analyzing the linguistic features of conspiracy theories, NLP models can enhance the accuracy of misinformation detection systems. The integration of emotion analysis, event relation graphs, and multilingual models offers a comprehensive approach to identifying and understanding conspiracy theories across diverse platforms and languages. These advancements pave the way for more robust and adaptive misinformation detection tools, aiding policymakers, social media platforms, and researchers in mitigating the spread of harmful conspiratorial content. The interdisciplinary nature of this research underscores the importance of a hybrid collaboration between computational linguistics and social sciences to address the multifaceted challenges posed by digital misinformation.

6 Multilingual approaches to computational discourse analysis

This section deals with multilingual approaches to computational discourse analysis developed in the past years. Both multilingual NLP and discourse analysis are incredibly broad topics of study. These topics include a wide range of tasks, approaches, data types, and applications. An exhaustive overview is therefore impossible. Instead, this chapter discusses multilingual approaches to four different discourse-related topics of study: rhetorical analysis, measuring discourse cohesion, and topic modeling. These subtopics are chosen to reflect the diversity of approaches and objectives of multilingual NLP, but are not claimed to be exhaustive or perfectly representative. Each subsection discusses approaches, resources, and open challenges in order to provide the reader with a global overview of the state of the field so far.

6.1 Multilingual NLP for rhetorical analysis

This subsection deals with NLP for the classification of relationships between parts of the texts, such as paragraphs or sentences. I will only discuss multilingual approaches to this task; see section 2 for a theoretical background on this topic and monolingual approaches.

It should be noted that the generalizability of rhetorical structure analysis across languages is not clear-cut even when done manually. Iruskieta et al. (2015) contrast parallel texts in English, Spanish, and Basque using RST. Different languages have different strategies to express rhetorical relations, meaning that the exact same passage in different languages might contain different rhetorical relations. Moreover, different translation strategies can also lead to different rhetorical structures. Although the authors do not discuss computational multilingual RST, their findings showcase the difficulty of automating this type of analysis.

Arguably the biggest obstacle for the development of multilingual RST parsing systems is the lack of multilingual training data (Peng et al., 2022). Ideally, multilingual training data for RST consists of corpora with comparable types of text data annotated with the exact same annotation scheme for different languages. Although this exist (see for example Cao et al. (2018) and Peng et al. (2022)), this is only available for few languages. Moreover, databases often consist of only few examples even for English, with the available resources for other languages being even smaller (Liu et al., 2020).

Braud et al. (2017) aim to overcome the problem of incompatible treebanks by harmonizing corpora in nine different languages. Although they manage to harmonize the data type and annotation schemes, the corpora are not parallel, making it difficult to do an in-depth analyzes of cross-linguistic differences in rhetorical structure and rhetorical structure parsing. The authors use a neural model for the construction of rhetorical trees based on some input text. They also implement a multilingual model based on word features created with a bilingual dictionary.

Joint learning/transfer learning can actually be a way to overcome the shortage of data; turn multilinguality into a strength rather than an obstacle (Liu et al., 2020). However, multilingual data seems only useful when it is high quality and a lot: Iruskieta et al. (2015) develop a monolingual RST parser for the Basque language, which they find works better than using lots of multilingual data; however, adding a bit of multilingual data to a large monolingual dataset in the target lan-

guage seemed to improve performance a bit. Language proximity also seems to be an important factor here.

Liu et al. (2020) also focus on multilingual rhetoric structure theory parsing. They compare two approaches: multilingual vector representations and automatic translation. For the translation, they automatically translated the original EDU units (preserving the original segmentation). They trained an encoder-decoder model on both and compared the results for different languages. Both reached SOTA performance. Liu et al. (2020) is an example of a system that assumes given EDU boundaries for both training and testing. However, in practical applications, this is not at all a given, and automatically detecting these boundaries is not trivial. Muller et al. (2019) focus on the automated detection of discourse unit boundaries. Depending on the annotation scheme, these boundaries can be anywhere, or only on sentence boundaries. They used a bi-LSTM and compared ELMo and BERT embeddings to rule-based and BOW-based baselines. They found that BERT-based embeddings outperform all other models by a large margin, reaching F1s of over .9.

The same authors from Liu et al. (2020) present a more advanced neural pipeline for multilingual RST classification in Liu et al. (2021b). They take RST tree banks in different languages as input data and perform label harmonization and cross translation as preprocessing steps. Whereas Liu et al. (2020) used translation to a single language (English) as a way to align data from originally multilingual sources, Liu et al. (2021b) use cross-translation (to all languages in the corpus) as a data augmentation strategy. This data is then fed to a model that segments the text into EDU and transforms them into contextualized EDU representations, including boundary embeddings (similar to Shi et al. (2016)). These are then organized in a tree structure of rhetorical relations using a transformer-based classifier. Their model aims to minimize three losses at the time: one for EDU segmentation on the document level, one for the parsing of the tree structure, and one for the labels and nuclearity prediction. They report outperforming even monolingual RST parsers for English, showing that multilinguality can be a strength, rather than an obstacle. They also tested their model in a zero-shot setting (i.e. testing it on languages not present in the training data). Although this (expectedly) led to a drop in scores, the results were still acceptable.

Peng et al. (2022) use the model developed by Liu et al. (2021b) to jointly train an RST model on Chinese and English. They found that joint training of Chinese and English data, represented in a multilingual embedding space, performed better than monolingual training for Chinese. This was not the case for English; the authors hypothesize that this is due to there being enough monolingual data for English, with the multilingual setting only 'confusing' the model.

Wang et al. (2022) compare ontologies of rhetorical figures in different languages for the recognition of rhetorical and stylistic figures. The difficulty with using ontologies for multilingual rhetorical analysis is that they are incompatible in terms of annotation, terminology, definition of stylistic figures and rhetorical operations, and data structure. The authors systematically identify these issues and synthesize the ontologies in one big resource for Serbian, German, and English.

A lot of recent work in multilingual NLP for discourse analysis makes use of multilingual contextualized word embeddings. These are word embeddings that are projected in a multilingual space in such a way that semantically similar concepts in different language will cluster together. Godunova and Voloshina (2024) probe multilingual LLM's (XLM-RoBERTa, mBERT, mGPT, and mT5) for their knowledge on discourse information. They do not focus on RST in and of itself;

rather, they use RST and UD to group their battery of tests into meaningful groups based on the type of manipulation used. They found no clear difference in performance between low- and high resource languages and conclude that LLMs are actually capable of learning language-agnostic discourse structures. They also find that syntactic and pragmatic particularities of a language (mainly the way languages organize clauses and how they mark topics) play an important role in what the models have learned in terms of discourse.

6.2 Multilingual approaches to measuring discourse coherence

Cohesion and coherence are important metrics to measure the quality of a piece of discourse. Although they are often used interchangeably in everyday discourse, they are not the same. Lapshinova-Koltunski and Kunz (2014) defines coherence as referring to 'the cognitive aspects of establishing meaning relations during text processing' (p. 57) and cohesion as being about the usage of 'implicit linguistic means that signal how clauses and sentences are linked together to function as a whole' (p. 57). Which linguistic means a speaker has to their disposal in order to make their discourse more cohesive is largely language-dependent. Cross-lingual studies of discourse cohesion therefore warrant higher-level categories of discourse markers, in order to produce a representation of discourse that allows comparison Lapshinova-Koltunski and Kunz (2014). Other scholars focus on *topic*, rather than connectives, when assessing coherence: Zhao et al. (2022) defines coherence as 'the continuity of semantics in text'.

The most classic approach to measuring coherence is the order discrimination task, where a model is tested on its ability to distinguish between an original text and a version with shuffled sentences. The main advantage of this test is that it is very easy to generate data; the downside is that this type of 'shuffled-sentence' coherence is not very similar to naturally occurring types of incoherence, which makes it difficult to gauge how effective these models are in practical applications of coherence measuring such as readability assessment (Bengoetxea and Gonzalez-Dios, 2021), writing quality (Rama and Vajjala, 2021) or document summarization (Cioaca et al., 2020). Pishdad et al. (2020) propose a battery of test for coherence modelling and test neural models of coherence based on their tasks, while still avoiding the use of human raters. They propose systematically switching topics and/or connectives, mixing of documents, and different cloze tasks (i.e. completion tasks).

A large part of the coherency is about finding a metric that best corresponds to human judgments. Zhao et al. (2022) find that BERT-based metrics often do not correspond to human metrics. Moreover, metrics are often sensitive to the way the problem is formulated, making them difficult to use for comparison of models across datasets (Zhao et al., 2021).

A big multilingual corpus for writing quality assessment is Merlin Boyd et al. (2014). This corpus contains texts annotated with language proficiency scores of their non-native authors in several dimensions, including coherency. In this corpus, coherency is defined as the ability of a writer to use connective words to make the text flow well. Rama and Vajjala (2021) simply fine-tune multilingual embeddings from Laser and mBERT on this corpus and see how well it does. They find that a multilingual model is especially useful for under-resourced languages, once again underlining that the transfer training paradigm has turned multilingualism from an obstacle into a strength.

Another interesting application of measuring coherence is seeing to what extents LLMs are able to model discourse. Brunato et al. (2023) test XLM-Roberta on its ability to predict whether a target sentence is the original continuation of a prompt sentence. They found that the model did not perform well at cross-domain generalization, but it showcased a surprising multilingual transfer. This is in line with earlier results by Godunova and Voloshina (2024) (discussed earlier), who found that LLMs appear to be good at multilingual transfer learning of discourse structures.

An interesting practical application of perplexity, one of the most commonly used measures of semantic coherence, is explored by Colla et al. (2022). They measure the perplexity of different LLMs when confronted with speech of disordered and healthy people, and find that the perplexity of all models was significantly higher when confronted with disordered speech, resulting in an *F1* of 1.00.

6.3 Multilingual approaches to topic modeling

Topic modeling is the task of discovering recurring topics in a corpus of documents. These topics are supposed to be semantically unified themes that are easily interpretable for humans. Topic modeling algorithms are typically applied in situations where a researchers wants to get an overview of the discourse topics in a large dataset. For example, (Mohawesh et al., 2023) use topic modeling for multilingual fake news detection. They use topic modeling to detect latent topics in fake and real news articles, which they use as the input for a graph neural network architecture. Malaterre and Lareau (2022) use it to gain insight in what types of topics were discussed by philosophers of science in the 20th century. They use machine translated versions of texts in various European languages and compare the different topics in order to gain a complete image in how the field of philosophy of science progressed in its early days. The multilingual topic modeling approach allows them to track how the research agenda evolves throughout time and linguistic/cultural spaces.

The most commonly used algorithm for topic modeling is Latent Dirichlet analysis (LDA), developed by Blei et al. (2003) for collections of discrete data. The core idea is that the distribution of discrete variables is generated by some underlying (*latent*) parameter. In the case of topic modeling, this means that the distribution of words over documents in a corpus is generated by an underlying distribution of topics over documents. The goal of LDA is to find these topics and the words associated with each topic by assigning distributions of topics to words. This means that each document in the corpus can be associated with multiple topics.

The question of how to do multilingual topic modeling is not easily answered. Classic LDA is based on the assumption of the data being *discrete*, i.e. the model treats all words as discrete categories, as opposed to models that treat words as meaningful projections in a shared vector space. Just giving a model input in different languages will thus result in 1) topics that are completely separated by language and 2) a vocabulary so large that it will make the model very difficult to train and fit (Boyd-Graber et al., 2014). This means that there needs to be a mapping of words from language A to language B; however, such one on one mappings are not easily available and not always possible to make. Moreover, topic modeling relies on word (or lemma) distribution. Different languages not only have different words, they also have different word distributions.

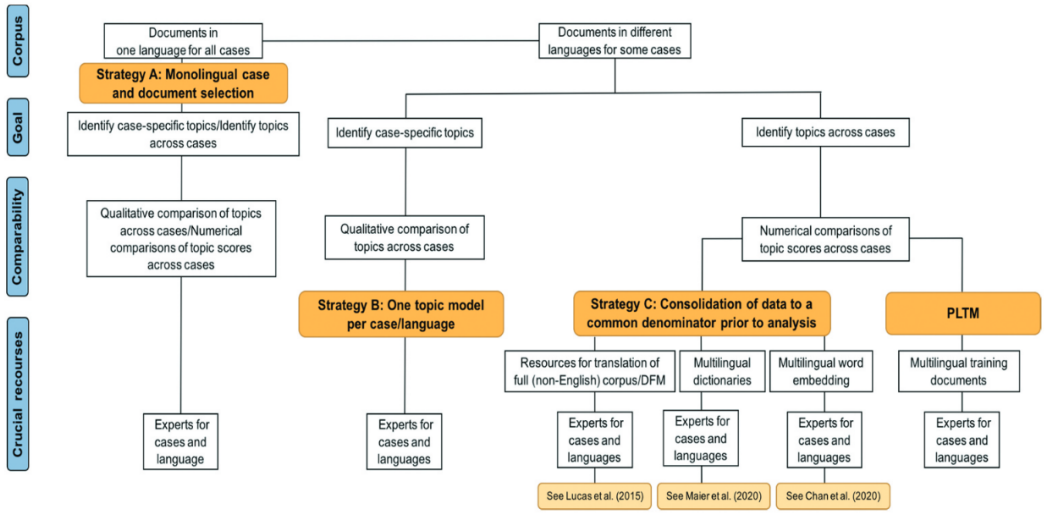


Figure 12: Taxonomy of approaches to topic modeling. Image by Lind et al. (2022), p. 98

There exist several strategies to multilingual topic modeling. Figure 12 shows a taxonomy of approaches by Lind et al. (2022). A very simple solution is to just run different topic modeling models for different language and then compare them manually (Heidenreich et al., 2019; Amara et al., 2021). Although this can be a feasible approach for research that works with manual/qualitative analysis of the corpus studied, I will not consider these works in this section as they do not involve multilingual NLP algorithms.

Multilingual approaches that treat documents as sets of discrete variables often use some kind of mapping strategy, where words, sentences, or documents are mapped to one semantic space (discrete or continuous) before the actual modeling process. Dictionary strategies make use of some kind of structure that maps words in language A to words in language B. This type of model often performs poorly, partially because of the different word distributions across languages. Shi et al. (2016) implement a dictionary mapping model for cross-cultural topic detection in Chinese and English texts, but improve it by adding an auxiliary word distribution (corresponding to the 'neutral' word distribution in a given language) which acts as a covariate. This decouples the topic distribution (terms typical for a topic) from the language-specific word distribution.

A recent example of the usage of parallel dictionaries for multilingual topic modeling is the work by Maier et al. (2022), who compare machine translation to a semantic-coding approach based on a multilingual dictionary. They study newspaper data and Twitter discourse in Arabic, English, and Hebrew on the conflict between Israeli settlers, Palestinians from the Westbank, and Israeli authorities. For the semantic-coding approach they used a multilingual dictionary that merges several words into larger concepts. This dictionary focuses explicitly on conflict situations. Interestingly, they found no structural differences between the two approaches when applied to journalistic articles, but they did when applying them to the Twitter data. Here, the topics created with the multilingual dictionary were found to be more specific and nuanced, whereas the topics created by the machine translation approach were found to be more broad and general.

Another type of mapping is based on parallel corpora, where some document in language A is

linked to its translation in language B. A very early paper using this approach was the Polylingual Topic Model (PLTM) by Mimno et al. (2009). They use a parallel corpus and assume that linked documents have the same distribution of topics. The algorithm then needs to find a topic distribution over tuples of linked document. Each topic consists of a set of distributions over words, where each set corresponds to a language. The authors propose 'glue documents' for cases in which no fully parallel corpus is available. In this case, tuples of parallel articles in different languages are used to align the topic models, whereas single-element tuples are used to improve language-specific topic distributions.

A problem with dictionary methods and parallel corpora is that these datasets are not always readily available. Creating parallel dictionaries is expensive, slow, and often imprecise. And although there are some large parallel corpora (such as Europarl or corpora of subtitles), the types of data these are available for is rather limited. Some researchers aim to solve it by using *comparable corpora* (Vulić et al., 2015). These are collections of documents that are matched with documents in another language on the base of discussing the same topic. Comparable corpora can be used for multilingual probabilistic topic modeling (Vulić et al., 2015). This approach aims to find *latent cross-lingual topics* that have language-specific representations given by per-topic word distributions for each language. This model needs to be trained on a topic-matched corpus (or *comparable corpus*) and can then be applied to unseen data in any of the languages it has been trained on. The most poignant example of a comparative corpus is Wikipedia. Another option is to sample for example newspaper articles with the same time stamps that mention the same entities; it can safely be assumed that these refer to broadly the same events and topics.

Piccardi and West (2021) also use Wikipedia as an alignment strategy. But rather than seeing the same article in different languages as equivalent bags of words, they treat them as equivalent bags of *links*. The advantage of this method is that these links represent language-independent concepts. So first they map the monolingual documents (i.e. the Wikipedia articles) to language-agnostic bags-of-links, and then they use this as the input for a topic model. They use LDA, but it could be the input for any type of topic model.

With the development of large language models and multilingual word embeddings Conneau et al. (2020), researchers have implemented strategies to use these as a base for topic modeling. One approach is to adjust the classic LDA model in order for it to allow for continuous (non-discrete) input. Das et al. (2015) introduce Gaussian LDA, a model that takes embeddings as input and defines topics as Gaussian distributions of probabilities over the multidimensional input space. A hybrid approach was taken by Sia and Duh (2021), who introduce a model that is a mixture of continuous (embedding) and discrete (word co-occurrence) representations. They use an adaptive coefficient that gauges the reliability of each type of feature and places more weight on the type it estimates to be more reliable. The authors found this approach to be especially useful in settings where the multilingual embeddings were of low quality.

Sia et al. (2020) compare LDA-based approaches to clustering of word embeddings. They extract contextualized embeddings and average them for each word type. They then performed a weighted centroid-based clustering, using the LDA-based definition of the probability of a word belonging to a topic to provide the weights. They found that, although their method does not outperform embedding-based LDA, it does yield equivalent results, with less computational complexity and shorter running time.

Xie et al. (2020) use BERT sentence embeddings for an LDA analysis to detect topics in scientific publications written in Chinese and English. They performed the topic modeling analysis separately for the two languages and then compared and matched topics by calculating the distances between the averaged embeddings

Chang and Hwang (2021) use multilingual word embeddings as the input for a model called center-based cross-lingual topic model. They define their topics as vectors in the multilingual embedding space. The probability of a word belonging to a topic is then calculated as a function of its distance to this topic vector. This resulted in topics that were clustered by language, which the authors solved by transforming the multilingual embeddings before the actual modeling, removing dimensions that have too high of a predictive power for what language it comes from.

A simple approach to contextualized embeddings for topic modeling is used by Zhang et al. (2022), who simply cluster contextualized sentence embeddings in the search for topics. A big difference from classic topic modeling is that it clusters whole *documents*, rather than finding a distribution of topics over documents. Although their experiments only use English data, this approach could be extended and adapted to multilingual sentence embeddings.

Recently, several researchers have explored (and successfully implemented) neural topic models (see Zhao et al. (2021) for a survey). An often-used base for neural topic model is the model introduced by Miao et al. (2016). They use two modules. One module aims to regenerate the words in a document based on its embedding; the second aims to predict whether a pair of a question and an answer is correct (the answer answers the question) or not. Although their model takes discrete documents as an input, it can be modified to allow for multilingual document embeddings as an input. Srivastava and Sutton (2017) use variational inference for topic modeling.

Neural topic models are easily scaled to more complex models and allow for better integration and joint training with other neural models. They also allow for the incorporation of metadata. Moreover, neural models can take word embeddings as input. This allows for the use of pretrained multilingual embeddings as input for the topic model. This, however, opens up a new problem, namely how to infer *interpretable* topics. This is often solved by adding a module that re-generates words based on the observed latent variables (Miao et al., 2016; Srivastava and Sutton, 2017; Bianchi et al., 2021).

Bianchi et al. (2021) create a zero-shot multilingual topic model, trained on English and tested on Italian, French, German, and Portuguese. The advantage of this model is that it is trained *only* on English, and the original English topics are then forced on the unseen languages, thus avoiding the problem of separation by language.

Many studies find that fine-tuning can significantly improve the usefulness of pretrained embeddings. But it is not immediately obvious how one can fine-tune a model for an unsupervised task. Mueller and Dredze (2021) experiment with several fine-tuning strategies for both monolingual and zero-shot multilingual neural topic modeling. They find that any kind of fine-tuning (no matter the task) improves the quality of the multilingual alignment of the topics. However, although their model produces good results for each individual language, this cross-lingual alignment remains poor.

Another problem is that many cross-lingual topic are of poor quality; they are not very different and tend to be repetitive (Mueller and Dredze, 2021). Wu et al. (2023) aim to resolve this by

maximizing distance between topics. Whereas basic topic models only maximize similarity within the topic, they also aim to maximize dissimilarity between topics, resulting in better and more informative topics. They also do a type of broad dictionary linking; they link a word not only to its direct dictionary translation, but also to words with embeddings that are close by in the monolingual embedding space of the target language. This also enables more interesting cross-lingual topics.

6.3.1 Evaluation

Topic models are supposed to extract semantically coherent topics from text. These topics can be rated by human annotators in terms of semantic coherence, topic diversity, and document coverage. However, this does not allow for easy comparison between models, and it is relatively expensive and subject to subtle differences in rater guidelines. Lau et al. (2014) use the manual metrics of word intrusion (meaning the relative amount of annotators who notice the 'intruder word' in a topic) and coherence (the 'coherence rating' given by human annotators). They propose a method to automize these methods using log likelihoods of synthetically generated word intruders.

Several automatized metrics have been proposed for the evaluation of topic models. Perplexity measures the 'surprisedness' of a model when seeing a certain piece of data. A topic model's perplexity over held-out data can be seen as a metric for how well the model is able to predict the topic distribution over data. Howevewr, this metric does not generally correlate with human judgments, and it is not concerned with topic quality (Zhao et al., 2021).

Topic coherence and topic diversity can be used to assess the quality of the produced topics. These metrics tend to be sensitive to small difference in formulation of the topic; Zhao et al. (2021) therefore recommend to calculate different metrics and take the average as a general indicator of topic quality in a model. A problem with these measures, especially when applied to multilingual topic models, is that they focus on intra-topic properties, and not on inter-topic properties, even though those tend to be the most problematic for multilingual models: the problem is often that topics are repetitive, or separated by language rather than topic. These metrics are not able to adequately capture them. A solution is proposed by Nan et al. (2019), who use *topic uniqueness* as an additional metric to measure topic model quality. Chang and Hwang (2021) combine metrics of topic diversity and coherence with a crosslingual-specific metric. For their parallel corpus, they calculate the convergence of assigned topics between parallel sentences. For their non-parallel corpus, they introduce a probing-based metric that uses the predicted topics in language A to train a classifier that is tested on data from language B.

Bianchi et al. (2021) evaluate their zero shot model by comparing the results for the unseen languages to the results on the same data but machine translated. They compare the amount of matches (same topic predicted), centroid embeddings (of the five words describing the topics for English and non English), and topic distribution over English vs. non-English documents. They also use a qualitative (manual) evaluation.

6.3.2 Challenges and future work

A specific problem mentioned by several authors (Maier et al., 2022; Shi et al., 2016) is the noise added by different transliterations or translations of names. A possible solution could be some kind of entity linking pre-processing step.

To the best of my knowledge, there is currently no benchmark dataset for multilingual topic modeling. Moreover, different authors use a wide variety of metrics in order to estimate model performance. This makes it difficult to compare models. On the other hand, just having a benchmark dataset would not solve this problem: there is something inherently subjective about the detection of good topics, and what exactly a model needs to do will depend on the specific data and application. Or, as said by Maier et al. (2022): 'We cannot say which approach is better, but instead focus on what each approach is better *at*' (p. 33)

6.4 Conclusion

As we have seen in this chapter, there are a wide range of approaches and applications for multilingual NLP in the domain of discourse analysis. Generally, the past 5 years have seen a shift to multilingual pretrained embeddings and transfer learning frameworks. Although these contextualized word representations are easily implemented in neural architectures, they are not so easily combined with models that presuppose a discrete BoW representation of text, such as classic LDA.

A common problem across multilingual tasks is the lack of high quality datasets. Data availability is especially a problem for RST analysis, where annotation is difficult and expensive. Other fields also deal with this problem, but several strategies have been developed in order to overcome the lack of ideal datasets.

The last problem is evaluation. Discourse analysis tasks are often complex and subjective, meaning the performance of a model is usually not well captured by one single number. On the other hand, the absence of a commonly adopted metric makes it difficult for users and other scientists to compare models and/or decide which one will best fit their use case.

7 Overall Conclusion

This document represents the in-depth work in systematically addressing the different scientific approaches currently existing within the computational treatment of discourse, both at a formal, methodological, and algorithmic level.

The D1.2 report encompasses the work done by five DCs in the doctoral network (DC 1 – 5), i.e. those that are mainly focused on the application of human and social sciences and in a more discourse-oriented linguistic level. It is complementary to the D1.1 Technical report on the state of the art on hybrid methods in NLP, in which more algorithmic works related to other linguistic levels are summarized. Thus, both deliverables fulfilled the WP1 aim which is designing the technological objectives based on Hybrid Intelligence (RO1). WP1 encompasses all doctoral researchers (DCs) as they will all follow the HI strategy which is the methodological pillar that supports the project.

We cover in a transversal way two main aspects of the relationship between discourse and its technological approaches. Firstly, we present the main formal theories for representing discourse computationally: RST (Section 2), Dialogical Argumentation (Section 3), Speech Act Theory (Section 4), and ad hoc formalizations depending on the domain, focusing on conspiracy theories as an illustrative example (Section 4). Secondly, we review existing algorithms and methods for several computational tasks related to discourse in a multilingual sphere (Section 5), including parsing, topic modeling, discourse analysis, or coherence measurement.

Thus, this document represents the theoretical discourse formalisms and the up-to-date algorithms upon which HYBRIDS DCs' activities will be based, and that will inform the implementation of new tools and approaches for a hybrid intelligence discourse treatment.

References

Abbott, R., Ecker, B., Anand, P., and Walker, M. (2016). Internet argument corpus 2.0: An sql schema for dialogic social media and the corpora to go with it. In Chair), N. C. C., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Abdelnabi, S., Hasan, R., and Fritz, M. (2022). Open-domain, content-based, multi-modal fact-checking of out-of-context images via online resources. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14920–14929.

Ahmed, H., Traore, I., and Saad, S. (2017). Detection of online fake news using n-gram analysis and machine learning techniques. In Traore, I., Woungang, I., and Awad, A., editors, *Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments*, volume 10618 of *Lecture Notes in Computer Science*, pages 127–138, Cham. Springer.

Akhtar, M., Cocarascu, O., and Simperl, E. (2022). Pubhealthtab: A public health table-based dataset for evidence-based fact checking. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1–16.

ALDayel, A. and Magdy, W. (2021). Stance detection on social media: State of the art and trends. *Information Processing & Management*, 58(4):102597.

Alhindi, T., Muresan, S., and Preotiuc-Pietro, D. (2020). Fact vs. Opinion: The Role of Argumentation Features in News Classification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6139–6149, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Alturayeif, N., Luqman, H., and Ahmed, M. (2023). A systematic review of machine learning techniques for stance detection and its applications. *Neural Computing and Applications*, 35(7):5113–5144.

Álvarez-Peralta, M., Rojas-Andrés, R., and Diefenbacher, S. (2023). Meta-analysis of political communication research on Twitter: Methodological trends. *Cogent Social Sciences*, 9(1):2209371.

Amara, A., Hadj Taieb, M. A., and Ben Aouicha, M. (2021). Multilingual topic modeling for tracking covid-19 trends based on facebook data analysis. *Applied Intelligence*, 51:3052–3073.

Anastasiou, L. and De Libbo, A. (2023). BCause: Reducing group bias and promoting cohesive discussion in online deliberation processes through a simple and engaging online deliberation tool. In *Proceedings of the First Workshop on Social Influence in Conversations (SICon 2023)*, pages 39–49, Toronto, Canada. Association for Computational Linguistics.

Andone, C. (2008). Douglas Walton, Dialog Theory for Critical Argumentation. *Argumentation*, 22(2):291–296.

Antypas, D., Camacho-Collados, J., Preece, A., and Rogers, D. (2021). COVID-19 and misinformation: A large-scale lexical analysis on Twitter. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 119–126, Online. Association for Computational Linguistics.

Arora, S., Rana, A., and Singh, A. (2023). Argument Mining: A Categorical Review. In Agrawal, R., Kishore Singh, C., Goyal, A., and Singh, D. K., editors, *Modern Electronics Devices and Communication Systems*, volume 948, pages 353–367. Springer Nature Singapore, Singapore.

Asterhan, C. S. C. and Schwarz, B. B. (2009). Argumentation and Explanation in Conceptual Change: Indications From Protocol Analyses of Peer-to-Peer Dialog. *Cognitive Science*, 33(3):374–400.

Atkinson, K., Bench-Capon, T., and McBurney, P. (2006). PARMENIDES: Facilitating Deliberation in Democracies. *Artificial Intelligence and Law*, 14(4):261–275.

Austin, J. L. (1975). *How to do things with words*. Harvard university press.

Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W., et al. (2023). A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Bansal, G., Chamola, V., Hussain, A., Guizani, M., and Niyato, D. (2024). Transforming Conversations with AI—A Comprehensive Study of ChatGPT. *Cognitive Computation*.

Barbedette, A. and Eshkol-Taravella, I. (2020). What speakers really mean when they ask questions: Classification of intentions with a supervised approach. *LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings*, pages 1159–1166.

Baturo, A., Dasandi, N., and Mikhaylov, S. J. (2017). Understanding state preferences with text as data: Introducing the un general debate corpus. *Research & Politics*, 4(2):2053168017712821.

Belcastro, L., Cantini, R., Marozzo, F., Talia, D., and Trunfio, P. (2020). Learning Political Polarization on Social Media Using Neural Networks. *IEEE Access*, 8:47177–47187.

Bengoetxea, K. and Gonzalez-Dios, I. (2021). Multiaztertest: A multilingual analyzer on multiple levels of language for readability assessment. *arXiv preprint arXiv:2109.04870*.

Bian, T., Xiao, X., Xu, T., Zhao, P., Huang, W., Rong, Y., and Huang, J. (2020). Rumor detection on social media with bi-directional graph convolutional networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 549–556.

Bianchi, F., Terragni, S., Hovy, D., Nozza, D., and Fersini, E. (2021). Cross-lingual contextualized topic models with zero-shot learning. In Merlo, P., Tiedemann, J., and Tsarfaty, R., editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1676–1683, Online. Association for Computational Linguistics.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Boltuzic, F. and Šnajder, J. (2016). Fill the Gap! Analyzing Implicit Premises between Claims from Online Debates. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 124–133, Berlin, Germany. Association for Computational Linguistics.

Boyd, A., Hana, J., Nicolas, L., Meurers, D., Wisniewski, K., Abel, A., Schöne, K., Stindlová, B., and Vettori, C. (2014). The merlin corpus: Learner language and the cefr. In *LREC*, pages 1281–1288. Reykjavik, Iceland.

Boyd-Graber, J., Mimno, D., and Newman, D. (2014). Care and feeding of topic models: Problems, diagnostics, and improvements. *Handbook of mixed membership models and their applications*, 225255.

Braud, C., Coavoux, M., and Søgaard, A. (2017). Cross-lingual RST discourse parsing. In Lapata, M., Blunsom, P., and Koller, A., editors, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 292–304, Valencia, Spain. Association for Computational Linguistics.

Brotherton, R. (2015). *Suspicious minds: Why we believe conspiracy theories*. Bloomsbury Publishing.

Brunato, D., Dell'Orletta, F., Dini, I., and Ravelli, A. A. (2023). Coherent or not? stressing a neural language model for discourse coherence in multiple languages. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10690–10700.

Buchholz, M. G. (2023). Assessing the effectiveness of gpt-3 in detecting false political statements: A case study on the liar dataset. *arXiv preprint arXiv:2306.08190*.

Budzynska, K. and Reed, C. (2011). Speech Acts of Argumentation: Inference Anchors and Peripheral Cues in Dialogue. In *AAAI Workshop - Technical Report*.

Budzynska, K. and Reed, C. (2019). Advances in Argument Mining. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 39–42, Florence, Italy. Association for Computational Linguistics.

Bunt, H., Petukhova, V., Gilmartin, E., Pelachaud, C., Fang, A., Keizer, S., and Prévot, L. (2020). The iso standard for dialogue act annotation. In *12th Edition of its Language Resources and Evaluation Conference (LREC 2020)*, pages 549–558. European Language Resources Association (ELRA).

Cambridge Dictionary (2024). Deontic. https://dictionary.cambridge.org/dictionary/english/deontic. Accessed: 2024-05-03.

Cao, S., da Cunha, I., and Iruskieta, M. (2018). The RST Spanish-Chinese treebank. In Savary, A., Ramisch, C., Hwang, J. D., Schneider, N., Andresen, M., Pradhan, S., and Petruck, M. R. L., editors, *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 156–166, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Caramancion, K. M. (2023). Harnessing the power of chatgpt to decimate mis/disinformation: Using chatgpt for fake news detection. In *2023 IEEE World AI IoT Congress (AIIoT)*, pages 0042–0046. IEEE.

Carlson, L., Marcu, D., and Okurovsky, M. E. (2001). Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*.

Casalegno, P., Iacona, A., Frascolla, P., Paganini, E., Santambrogio, M. L. V., et al. (2003). *Filosofia del linguaggio*. R. Cortina.

Chai, J. and Jin, R. (2004). Discourse structure for context question answering. In *Proceedings of the Workshop on Pragmatics of Question Answering at HLT-NAACL 2004*, pages 23–30.

Chang, C.-H. and Hwang, S.-Y. (2021). A word embedding-based approach to cross-lingual topic modeling. *Knowledge and Information Systems*, 63(6):1529–1555.

Chang, J. P. and Danescu-Niculescu-Mizil, C. (2019). Trouble on the Horizon: Forecasting the Derailment of Online Conversations as they Develop. In Inui, K., Jiang, J., Ng, V., and Wan, X., editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4743–4754, Hong Kong, China. Association for Computational Linguistics.

Chen, C. and Shu, K. (2023a). Can llm-generated misinformation be detected? *arXiv preprint arXiv:2309.13788*.

Chen, C. and Shu, K. (2023b). Combating misinformation in the age of llms: Opportunities and challenges. *arXiv preprint arXiv:2311.05656*.

Chen, J., Sriram, A., Choi, E., and Durrett, G. (2022). Generating literal and implied subquestions to fact-check complex claims. *arXiv preprint arXiv:2205.06938*.

Chen, W., Zhang, X., Wang, T., Yang, B., and Li, Y. (2017). Opinion-aware Knowledge Graph for Political Ideology Detection. pages 3647–3653.

Chernyavskiy, A., Ilvovsky, D., and Nakov, P. (2024). Unleashing the power of discourse-enhanced transformers for propaganda detection. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1452–1462.

Cheung, T.-H. and Lam, K.-M. (2023). Factllama: Optimizing instruction-following language models with external knowledge for automated fact-checking. In *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 846–853. IEEE.

Chong, D., Lee, E., Fan, M., Holur, P., Shahsavari, S., Tangherlini, T., and Roychowdhury, V. (2021). A real-time platform for contextualized conspiracy theory analysis. In *2021 International Conference on Data Mining Workshops (ICDMW)*, pages 118–127. IEEE.

Chowdhury, R., Srinivasan, S., and Getoor, L. (2020). Joint estimation of user and publisher credibility for fake news detection. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, page 1993–1996, New York, NY, USA. Association for Computing Machinery.

Chu, S. K. W., Xie, R., and Wang, Y. (2021). Cross-language fake news detection. *Data and Information Management*, 5(1):100–109.

Ciampaglia, G. L., Shiralkar, P., Rocha, L. M., Bollen, J., Menczer, F., and Flammini, A. (2015). Computational fact checking from knowledge networks. *PloS one*, 10(6):e0128193.

Cioaca, V. S., Dascalu, M., and McNamara, D. S. (2020). Extractive summarization using cohesion network analysis and submodular set functions. In *2020 22nd International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)*, pages 161–168. IEEE.

Cohen, W., Carvalho, V., and Mitchell, T. (2004). Learning to classify email into "speech acts". *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, EMNLP 2004 - A meeting of SIGDAT, a Special Interest Group of the ACL held in conjunction with ACL 2004*, pages 309–316.

Colla, D., Delsanto, M., Agosto, M., Vitiello, B., and Radicioni, D. P. (2022). Semantic coherence markers: The contribution of perplexity metrics. *Artificial Intelligence in Medicine*, 134:102393.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Conover, M. D., Goncalves, B., Ratkiewicz, J., Flammini, A., and Menczer, F. (2011). Predicting the Political Alignment of Twitter Users. In *2011 IEEE Third Int'l Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third Int'l Conference on Social Computing*, pages 192–199, Boston, MA, USA. IEEE.

Cui, C., Mao, W., Zheng, X., and Zeng, D. (2017). Mining user intents in online interactions: Applying to discussions about medical event on SinaWeibo platform. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 10347:177–183.

Cui, L., Seo, H., Tabar, M., Ma, F., Wang, S., and Lee, D. (2020). Deterrent: Knowledge guided graph attention network for detecting healthcare misinformation. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 492–502.

da Silva, N. L. P., Roman, N. T., and Felippo, A. D. (2024). Bringing Pragmatics to Porttinari - Adding Speech Acts to News Texts. In Gamallo, P., Claro, D., Teixeira, A., Real, L., Garcia, M., Oliveira, H. G., and Amaro, R., editors, *Proceedings of the 16th International Conference on Computational Processing of Portuguese*, pages 137–145, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics.

Das, D. and Taboada, M. (2018). Rst signalling corpus: a corpus of signals of coherence relations. *Language Resources and Evaluation*, 52.

Das, D. and Taboada, M. (2019). Multiple signals of coherence relations. *Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics*, (24).

Das, R., Zaheer, M., and Dyer, C. (2015). Gaussian lda for topic models with word embeddings. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 795–804.

Dascălu, M. and Dascalu, M. (2014). Computational discourse analysis. *Analyzing Discourse and Text Complexity for Learning and Collaborating: A Cognitive Approach Based on Natural Language Processing*, pages 53–77.

Del Tredici, M. and Fernández, R. (2020). Words are the window to the soul: Language-based user representations for fake news detection. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5467–5479, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Diaz, M., Amironesei, R., Weidinger, L., and Gabriel, I. (2022). Accounting for offensive speech as a practice of resistance. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 192–202, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Douglas, K. M. and Sutton, R. M. (2008). The hidden impact of conspiracy theories: Perceived and actual influence of theories surrounding the death of princess diana. *The Journal of social psychology*, 148(2):210–222.

Egan, C., Siddharthan, A., and Wyner, A. (2016). Summarising the points made in online political debates. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 134–143, Berlin, Germany. Association for Computational Linguistics.

Epure, E., Deneckere, R., and Salinesi, C. (2017). Analyzing perceived intentions of public health-related communication on Twitter. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 10259:182–192.

Fairclough, N. (1996). A reply to henry widdowson's 'discourse analysis: a critical view'. *Language and Literature*, 5(1):49–56.

Farinneya, P., Pour, M. M. A., Hamidian, S., and Diab, M. (2021). Active learning for rumor identification on social media. In *Findings of the association for computational linguistics: EMNLP 2021*, pages 4556–4565.

Felton, M., Crowell, A., Garcia-Mila, M., and Villarroel, C. (2022). Capturing deliberative argument: An analytic coding scheme for studying argumentative dialogue and its benefits for learning. *Learning, Culture and Social Interaction*, 36:100350.

Felton, M., Garcia-Mila, M., Villarroel, C., and Gilabert, S. (2015). Arguing collaboratively: Argumentative discourse types and their potential for knowledge building. *British Journal of Educational Psychology*, 85(3):372–386.

Fromkin, V., Rodman, R., Hyams, N., Amberber, M., Cox, F., and Thornton, R. (2017). *An Introduction to Language with Online Study Tools 12 Months*. Cengage AU.

Gearhart, S., Moe, A., and Zhang, B. (2020). Hostile media bias on social media: Testing the effect of user comments on perceptions of news bias and credibility. *Human Behavior and Emerging Technologies*, 2.

Ghafouri, V., Agarwal, V., Zhang, Y., Sastry, N., Such, J., and Suarez-Tangil, G. (2023). AI in the Gray: Exploring Moderation Policies in Dialogic Large Language Models vs. Human Answers in Controversial Topics. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 556–565, Birmingham United Kingdom. ACM.

Godden, D. and Wells, S. (2022). Burdens of Proposing: On the Burden of Proof in Deliberation Dialogues. *Informal Logic*, 42(1):291–342.

Godunova, M. and Voloshina, E. (2024). Probing of pretrained multilingual models on the knowledge of discourse. In *Proceedings of the 5th Workshop on Computational Approaches to Discourse (CODI 2024)*, pages 78–90.

Goffredo, P., Cabrio, E., Villata, S., Haddadan, S., and Torres Sanchez, J. (2023). DISPUTool 2.0: A Modular Architecture for Multi-Layer Argumentative Analysis of Political Debates. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 16431–16433.

Gordon, T. F., Prakken, H., and Walton, D. (2007). The Carneades model of argument and burden of proof. *Artificial Intelligence*, 171(10):875–896.

Gordon, T. F. and Walton, D. (2009). Proof Burdens and Standards. *Argumentation in Artificial Intelligence*, page 239.

Grice, H. P. (1969). Utterer's meaning and intention. *The philosophical review*, 78(2):147–177.

Guo, Z. and Singh, M. P. (2023). Representing and Determining Argumentative Relevance in Online Discussions: A General Approach. *Proceedings of the International AAAI Conference on Web and Social Media*, 17:292–302.

Gurcke, T., Alshomary, M., and Wachsmuth, H. (2021). Assessing the Sufficiency of Arguments through Conclusion Generation. In *Proceedings of the 8th Workshop on Argument Mining*, pages 67–77, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Halpern, D., Katz, J. E., and Carril, C. (2017). The online ideal persona vs. the jealousy effect: Two explanations of why selfies are associated with lower-quality romantic relationships. *Telematics and Informatics*, 34(1):114–123.

Hardalov, M., Arora, A., Nakov, P., and Augenstein, I. (2022). A Survey on Stance Detection for Mis- and Disinformation Identification.

Harris, D. W., Fogal, D., and Moss, M. (2018). Speech Acts: The Contemporary Theoretical Landscape. In Fogal, D., Harris, D. W., and Moss, M., editors, *New Work on Speech Acts*. Oxford University Press.

Harris, D. W. and McKinney, R. (2021). Speech-Act Theory: Social and Political Applications. In *The Routledge Handbook of Social and Political Philosophy of Language*. Routledge. Num Pages: 21.

Hashim, S. S. M. and Safwat, S. (2015). Speech acts in political speeches. *Journal of Modern Education Review*, 5(7):699–706.

Haupt, M. R., Chiu, M., Chang, J., Li, Z., Cuomo, R., and Mackey, T. K. (2023). Detecting nuance in conspiracy discourse: Advancing methods in infodemiology and communication science with machine learning and qualitative content coding. *Plos one*, 18(12):e0295414.

Hautli-Janisz, A., Budzynska, K., McKillop, C., Plüss, B., Gold, V., and Reed, C. (2022). Questions in argumentative dialogue. *Journal of Pragmatics*, 188:56–79.

He, Z., Li, C., Zhou, F., and Yang, Y. (2021). Rumor detection on social media with event augmentations. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 2020–2024.

Heft, A. and Buehling, K. (2022). Measuring the diffusion of conspiracy theories in digital information ecologies. *Convergence*, 28(4):940–961.

Heidenreich, T., Lind, F., Eberl, J.-M., and Boomgaarden, H. G. (2019). Media framing dynamics of the 'european refugee crisis': A comparative topic modelling approach. *Journal of Refugee Studies*, 32(Special_Issue_1):i172–i182.

Hernault, H., Prendinger, H., du Verle, D. A., and Ishizuka, M. (2010). Hilda: A discourse parser using support vector machine classification. *Dialogue & Discourse*, 1(3):1–33.

Hinton, M. and Wagemans, J. H. (2023). How persuasive is AI-generated argumentation? An analysis of the quality of an argumentative text produced by the GPT-3 AI text generator. *Argument & Computation*, 14(1):59–74.

Hitchcock, D., Mcburney, P., and Parsons, S. (2001). A Framework for Deliberation Dialogues. *OSSA Conference Archive*.

Hornikx, J. (2013). Een Bayesiaans perspectief op argumentkwaliteit - Het ad populum-argument onder de loep. *Tijdschrift voor Taalbeheersing*, 35(2):128–143.

Hou, S., Zhang, S., and Fei, C. (2020). Rhetorical structure theory: A comprehensive review of theory, parsing methods and applications. *Expert Systems with Applications*, 157:113421.

Hu, B., Sheng, Q., Cao, J., Zhu, Y., Wang, D., Wang, Z., and Jin, Z. (2023). Learn over past, evolve for future: Forecasting temporal trends for fake news detection. *arXiv preprint arXiv:2306.14728*.

Huang, Y., Gao, M., Wang, J., and Shu, K. (2021). Dafd: Domain adaptation framework for fake news detection. In *Neural Information Processing: 28th International Conference, ICONIP 2021, Sanur, Bali, Indonesia, December 8–12, 2021, Proceedings, Part I 28*, pages 305–316. Springer.

Huang, Y., Gao, M., Wang, J., Yin, J., Shu, K., Fan, Q., and Wen, J. (2023). Meta-prompt based learning for low-resource false information detection. *Information Processing & Management*, 60(3):103279.

Iandoli, L., Quinto, I., Spada, P., Klein, M., and Calabretta, R. (2018). Supporting argumentation in online political debate: Evidence from an experiment of collective deliberation. *New Media & Society*, 20(4):1320–1341.

Iruskieta, M., Da Cunha, I., and Taboada, M. (2015). A qualitative comparison method for rhetorical structures: identifying different discourse structures in multilingual corpora. *Language resources and evaluation*, 49:263–309.

ISO (2020). *ISO 24617-2:2020. Language resource management Semantic annotation framework*.

Janier, M., Lawrence, J., and Reed, C. (2014). OVA+: An Argument Analysis Interface. In *Computational Models of Argument*, pages 463–464. IOS Press.

Jin, M., Preotiuc-Pietro, D., Dogruöz, A., and Aletras, N. (2022). Automatic Identification and Classification of Bragging in Social Media. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 1:3945–3959.

Kalchbrenner, N., Grefenstette, E., and Blunsom, P. (2014). A convolutional neural network for modelling sentences. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 655–665, Baltimore, Maryland. Association for Computational Linguistics.

Kaliyar, R. K., Goswami, A., and Narang, P. (2021). Fakebert: Fake news detection in social media with a bert-based deep learning approach. *Multimedia tools and applications*, 80(8):11765–11788.

Kantartopoulos, P., Pitropakis, N., Mylonas, A., and Kylilis, N. (2020). Exploring Adversarial Attacks and Defences for Fake Twitter Account Detection. *Technologies*, 8(4):64.

Katzav, J. and Reed, C. A. (2004). On Argumentation Schemes and the Natural Classification of Arguments. *Argumentation*, 18(2):239–259.

Kawintiranon, K. and Singh, L. (2022). PoliBERTweet: A Pre-trained Language Model for Analyzing Political Content on Twitter. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Odijk, J., and Piperidis, S., editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7360–7367, Marseille, France. European Language Resources Association.

Kim, K., Kim, H., and Seo, J. (2004). A neural network model with feature selection for Korean speech act classification. *International journal of neural systems*, 14(6):407–414.

Kim, Y. and Allan, J. (2019). Unsupervised explainable controversy detection from online news. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11437:836–843.

Kobbe, J., Rehbein, I., Hulpu\textcommabelows, I., and Stuckenschmidt, H. (2020). Exploring Morality in Argumentation. In Cabrio, E. and Villata, S., editors, *Proceedings of the 7th Workshop on Argument Mining*, pages 30–40, Online. Association for Computational Linguistics.

Kok, E. M., Meyer, J.-J. C., Prakken, H., and Vreeswijk, G. A. W. (2011). A Formal Argumentation Framework for Deliberation Dialogues. In McBurney, P., Rahwan, I., and Parsons, S., editors, *Argumentation in Multi-Agent Systems*, volume 6614, pages 31–48. Springer Berlin Heidelberg, Berlin, Heidelberg.

Komalova, L., Glazkova, A., Morozov, D., Epifanov, R., Motovskikh, L., and Mayorova, E. (2022). Automated Classification of Potentially Insulting Speech Acts on Social Network Sites. *Communications in Computer and Information Science*, 1503:365–374.

Koo, Y., Kim, J., and Hong, M. (2019). Automatic speech act classification of Korean dialogue based on the hierarchical structure of speech act categories. *Proceedings of the 33rd Pacific Asia Conference on Language, Information and Computation, PACLIC 2019*, pages 432–441.

Kotonya, N. and Toni, F. (2019). Gradual Argumentation Evaluation for Stance Aggregation in Automated Fake News Detection. In *Proceedings of the 6th Workshop on Argument Mining*, pages 156–166, Florence, Italy. Association for Computational Linguistics.

Krabbe, E. C. W. (2002). Profiles of Dialogue as a Dialectical Tool. In van Eemeren, F. H., editor, *Advances in Pragma-Dialectics*, pages 153–167. Sic Sat & Vale Press, Amsterdam/Newport News, VA.

Kushwaha, A. K., Kar, A. K., Roy, S. K., and Ilavarasan, P. V. (2022). Capricious opinions: A study of polarization of social media groups. *Government Information Quarterly*, 39(3):101709.

Ladd, B. K. and Goodwin, J. (2022). Extreme arguments: Anwar al-Awlaki's radicalizing discourse. *Journal of Pragmatics*, 200:39–48.

Lai, M., Cignarella, A. T., Hernández Farías, D. I., Bosco, C., Patti, V., and Rosso, P. (2020). Multilingual stance detection in social media political debates. *Computer Speech & Language*, 63:101075.

Langguth, J., Schroeder, D. T., Filkuková, P., Brenner, S., Phillips, J., and Pogorelov, K. (2023). Coco: an annotated twitter dataset of covid-19 conspiracy theories. *Journal of Computational Social Science*, 6(2):443–484.

Lapshinova-Koltunski, E. and Kunz, K. (2014). Annotating cohesion for multilingual analysis. In *Proceedings 10th Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation*, pages 57–64.

Lau, J. H., Newman, D., and Baldwin, T. (2014). Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 530–539.

Laurenti, E., Bourgon, N., Benamara, F., Mari, A., Moriceau, V., and Courgeon, C. (2022). Give me your Intentions, I'll Predict our Actions: A Two-level Classification of Speech Acts for Crisis Management in Social Media. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Odijk, J., and Piperidis, S., editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4333–4343, Marseille, France. European Language Resources Association.

Lawrence, J., Janier, M., and Reed, C. (2016). Working with Open Argument Corpora: 1st European Conference on Argumentation. *Studies in Logic and Argumentation*, 1.

Lawrence, J. and Reed, C. (2017). Mining Argumentative Structure from Natural Language text using Automatically Generated Premise-Conclusion Topic Models. In *Proceedings of the 4th Workshop on Argument Mining*, pages 39–48, Copenhagen, Denmark. Association for Computational Linguistics.

Leech, G. N. (2016). *Principles of pragmatics*. Routledge.

Lei, Y. and Huang, R. (2023). Identifying conspiracy theories news based on event relation graph. *arXiv preprint arXiv:2310.18545*.

Lewiński, M. (2021). Illocutionary pluralism. *Synthese*, 199(3):6687–6714.

Li, J., Liu, M., Qin, B., and Liu, T. (2022a). A survey of discourse parsing. *Frontiers of Computer Science*, 16(5):165329.

Li, J., Sujana, Y., and Kao, H.-Y. (2020). Exploiting microblog conversation structures to detect rumors. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5420–5429, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Li, K., Guo, B., Ren, S., and Yu, Z. (2022b). Adadebunk: An efficient and reliable deep state space model for adaptive fake news early detection. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 1156–1165.

Li, M., Peng, B., and Zhang, Z. (2023). Self-checker: Plug-and-play modules for fact-checking with large language models. *arXiv preprint arXiv:2305.14623*.

Lind, F., Eberl, J.-M., Eisele, O., Heidenreich, T., Galyga, S., and Boomgaarden, H. G. (2022). Building the bridge: Topic modeling for comparative research. *Communication Methods and Measures*, 16(2):96–114.

Lippi, M. and Torroni, P. (2016). Argumentation Mining: State of the Art and Emerging Trends. *ACM Transactions on Internet Technology*, 16(2):1–25.

Lisanyuk, E. N. and Prokudin, D. E. (2021). Software for Modeling Deliberative Argumentation: Requirements and Criteria. *IMS 2021 - International Conference "Internet and Modern Society"*, pages 11–23.

Liu, Q., Wu, J., Wu, S., and Wang, L. (2024a). Out-of-distribution evidence-aware fake news detection via dual adversarial debiasing. *IEEE Transactions on Knowledge and Data Engineering*.

Liu, Y. J., Aoyama, T., and Zeldes, A. (2023). What's hard in English RST parsing? predictive models for error analysis. In *Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue*, pages 31–42, Prague, Czechia. Association for Computational Linguistics.

Liu, Z., Liu, B., Thompson, P., Yang, K., Jain, R., and Ananiadou, S. (2024b). Conspemollm: Conspiracy theory detection using an emotion-based large language model. *arXiv preprint arXiv:2403.06765*.

Liu, Z., Shi, K., and Chen, N. (2020). Multilingual neural RST discourse parsing. In Scott, D., Bel, N., and Zong, C., editors, *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6730–6738, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Liu, Z., Shi, K., and Chen, N. (2021a). DMRST: A joint framework for document-level multilingual RST discourse segmentation and parsing. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 154–164, Punta Cana, Dominican Republic and Online. Association for Computational Linguistics.

Liu, Z., Shi, K., and Chen, N. (2021b). DMRST: A joint framework for document-level multilingual RST discourse segmentation and parsing. In Braud, C., Hardmeier, C., Li, J. J., Louis, A., Strube, M., and Zeldes, A., editors, *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 154–164, Punta Cana, Dominican Republic and Online. Association for Computational Linguistics.

Lustick, I. S. and Miodownik, D. (2000). Deliberative democracy and public discourse: The agent-based argument repertoire model. *Complexity*, 5(4):13–30.

Lytos, A., Lagkas, T., Sarigiannidis, P., and Bontcheva, K. (2019). The evolution of argumentation mining: From models to social media and emerging tools. *Information Processing & Management*, 56(6):102055.

Macagno, F. (2021). Argumentation schemes in AI: A literature review. Introduction to the special issue. *Argument & Computation*, 12:1–16.

Macagno, F. (2022). Argumentation schemes, fallacies, and evidence in politicians' argumentative tweets—A coded dataset. *Data in Brief*, 44:108501.

Maier, D., Baden, C., Stoltenberg, D., De Vries-Kedem, M., and Waldherr, A. (2022). Machine translation vs. multilingual dictionaries assessing two strategies for the topic modeling of multilingual text collections. *Communication methods and measures*, 16(1):19–38.

Malaterre, C. and Lareau, F. (2022). The early days of contemporary philosophy of science: novel insights from machine translation and topic-modeling of non-parallel multilingual corpora. *Synthese*, 200(3):242.

Mann, W. C. and Thompson, S. A. (1987). *Rhetorical structure theory: A theory of text organization*. University of Southern California, Information Sciences Institute Los Angeles.

Marcellino, W., HELMUS, T. C., KERRIGAN, J., REININGER, H., KARIMOV, R. I., and LAWRENCE, R. A. (2021). Detecting conspirac y theories on social media.

Matley, D. (2018). "this is not a# humblebrag, this is just a# brag": The pragmatics of self-praise, hashtags and politeness in instagram posts. *Discourse, context & media*, 22:30–38.

McBurney, P., Rahwan, I., Simon, P., and Maudet, N., editors (2010). *Argumentation in Multi-Agent Systems*. Springer.

Mestre, R., Milicin, R., Middleton, S., Ryan, M., Zhu, J., and Norman, T. J. (2021). M-Arg: Multimodal Argument Mining Dataset for Political Debates with Audio and Transcripts. In *Proceedings of the 8th Workshop on Argument Mining*, pages 78–88, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Miani, A., Hills, T., and Bangerter, A. (2021). Loco: The 88-million-word language of conspiracy corpus. *Behavior research methods*, pages 1–24.

Miao, Y., Yu, L., and Blunsom, P. (2016). Neural variational inference for text processing. In *International conference on machine learning*, pages 1727–1736. PMLR.

Michikyan, M., Dennis, J., and Subrahmanyam, K. (2015). Can you guess who i am? real, ideal, and false self-presentation on facebook among emerging adults. *Emerging Adulthood*, 3(1):55–64.

Mimno, D., Wallach, H., Naradowsky, J., Smith, D. A., and McCallum, A. (2009). Polylingual topic models. In *Proceedings of the 2009 conference on empirical methods in natural language processing*, pages 880–889.

Misra, A., Anand, P., Tree, J. E. F., and Walker, M. (2015). Using Summarization to Discover Argument Facets in Online Ideological Dialog. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 430–440.

Misra, A., Oraby, S., Tandon, S., TS, S., Anand, P., and Walker, M. (2017). Summarizing Dialogic Arguments from Social Media.

Modgil, S. and Prakken, H. (2014). The ASPIC+ framework for structured argumentation: A tutorial. *Argument & Computation*, 5(1):31–62.

Mohawesh, R., Liu, X., Arini, H. M., Wu, Y., and Yin, H. (2023). Semantic graph based topic modelling framework for multilingual fake news detection. *AI Open*, 4:33–41.

Moldovan, C., Rus, V., and Graesser, A. (2011). Automated Speech Act Classification For Online Chat.

Mou, X., Li, Z., Lyu, H., Luo, J., and Wei, Z. (2024). Unifying Local and Global Knowledge: Empowering Large Language Models as Political Experts with Knowledge Graphs. In *Proceedings of the ACM on Web Conference 2024*, pages 2603–2614, Singapore Singapore. ACM.

Mueller, A. and Dredze, M. (2021). Fine-tuning encoders for improved monolingual and zero-shot polylingual neural topic modeling. In Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., and Zhou, Y., editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3054–3068, Online. Association for Computational Linguistics.

Muller, P., Braud, C., and Morey, M. (2019). Tony: Contextual embeddings for accurate multilingual discourse segmentation of full documents. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pages 115–124.

Myers West, S. (2018). Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society*, 20(11):4366–4383.

Nan, F., Ding, R., Nallapati, R., and Xiang, B. (2019). Topic modeling with wasserstein autoencoders. *arXiv preprint arXiv:1907.12374*.

Nguyen, H. and Gokhale, S. (2022). An efficient approach to identifying anti-government sentiment on Twitter during Michigan protests. *PeerJ Computer Science*, 8:e1127.

Nguyen, T.-T., Nguyen, X.-P., Joty, S., and Li, X. (2021). Rst parsing from scratch. *arXiv preprint arXiv:2105.10861*.

Nivre, J., de Marneffe, M.-C., Ginter, F., Hajič, J., Manning, C. D., Pyysalo, S., Schuster, S., Tyers, F., and Zeman, D. (2020). Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Nordenstam, T. (1966). On austin's theory of speech-acts. *Mind*, 75(297):141–143.

Pan, L., Wu, X., Lu, X., Luu, A. T., Wang, W. Y., Kan, M.-Y., and Nakov, P. (2023). Fact-checking complex claims with program-guided reasoning. *arXiv preprint arXiv:2305.12744*.

Pan, S., Luo, L., Wang, Y., Chen, C., Wang, J., and Wu, X. (2024). Unifying Large Language Models and Knowledge Graphs: A Roadmap. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–20.

Panchendrarajan, R. and Zubiaga, A. (2024). Synergizing machine learning & symbolic methods: A survey on hybrid approaches to natural language processing. *arXiv preprint arXiv:2401.11972*.

Parsons, C. (2007). *How to Map Arguments in Political Science*. Oxford University Press, Oxford ; New York.

Passon, M., Lippi, M., Serra, G., and Tasso, C. (2018). Predicting the Usefulness of Amazon Reviews Using Off-The-Shelf Argumentation Mining. In *Proceedings of the 5th Workshop on Argument Mining*, pages 35–39, Brussels, Belgium. Association for Computational Linguistics.

Pastor, M. and Oostdijk, N. (2024). Signals as features: Predicting error/success in rhetorical structure parsing. In Strube, M., Braud, C., Hardmeier, C., Li, J. J., Loaiciga, S., Zeldes, A., and Li, C., editors, *Proceedings of the 5th Workshop on Computational Approaches to Discourse (CODI 2024)*, pages 139–148, St. Julians, Malta. Association for Computational Linguistics.

Pastor, M., Oostdijk, N., and Larson, M. (2024). The contribution of coherence relations to understanding paratactic forms of communication in social media comment sections. In *JADT 2024: 17th International Conference on Statistical Analysis of Textual Data*.

Pelrine, K., Reksoprodjo, M., Gupta, C., Christoph, J., and Rabbany, R. (2023). Towards reliable misinformation mitigation: Generalization, uncertainty, and gpt-4. *arXiv preprint arXiv:2305.14928*.

Peng, S., Liu, Y. J., and Zeldes, A. (2022). GCDT: A Chinese RST treebank for multigenre and multilingual discourse parsing. In He, Y., Ji, H., Li, S., Liu, Y., and Chang, C.-H., editors, *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 382–391, Online only. Association for Computational Linguistics.

Peters, G. and Woolley, J. T. (2019). The state of the union, background and reference table. *The American Presidency Project*, pages 1999–2020.

Piccardi, T. and West, R. (2021). Crosslingual topic modeling with wikipda. In *Proceedings of the Web Conference 2021*, pages 3032–3041.

Pishdad, L., Fancellu, F., Zhang, R., and Fazly, A. (2020). How coherent are neural models of coherence? In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6126–6138.

Pitler, E., Raghupathy, M., Mehta, H., Nenkova, A., Lee, A., and Joshi, A. (2008). Easily identifiable discourse relations. In *Coling 2008: Companion volume: Posters*, pages 87–90, Manchester, UK. Coling 2008 Organizing Committee.

Platt, A., Brown, J., and Venske, A. (2022). Toward detecting conspiracy language in misinformation documents. In *Proceedings of the 2022 Computers and People Research Conference*, pages 1–4.

Prendinger, H., Piwek, P., and Ishizuka, M. (2007). A novel method for automatically generating multi-modal dialogue from text. *International Journal of Semantic Computing*, 1(03):319–334.

Pustet, M., Steffen, E., and Mihaljević, H. (2024). Detection of conspiracy theories beyond keyword bias in german-language telegram using large language models. *arXiv preprint arXiv:2404.17985*.

Rahwan, I., Banihashemi, B., Reed, C., Walton, D., and Abdallah, S. (2011). Representing and classifying arguments on the Semantic Web. *The Knowledge Engineering Review*, 26(4):487–511.

Rahwan, I. and Reed, C. (2009). The Argument Interchange Format. In Simari, G. and Rahwan, I., editors, *Argumentation in Artificial Intelligence*, pages 383–402. Springer US, Boston, MA.

Rama, T. and Vajjala, S. (2021). Are pre-trained text representations useful for multilingual and multi-dimensional language proficiency modeling? *arXiv preprint arXiv:2102.12971*.

Reid, C. E. (2011). Rationale Argument Mapping Software. *Journal of Technology in Human Services*, 29(2):147–154.

Reisert, P., Inoue, N., Okazaki, N., and Inui, K. (2015). A Computational Approach for Generating Toulmin Model Argumentation. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 45–55, Denver, CO. Association for Computational Linguistics.

Rigoli, F. (2022). Deconstructing the conspiratorial mind: The computational logic behind conspiracy theories. *Review of Philosophy and Psychology*, pages 1–18.

Rocha, G. and Lopes Cardoso, H. (2017). Towards a Relation-Based Argument Extraction Model for Argumentation Mining. In Camelin, N., Estève, Y., and Martín-Vide, C., editors, *Statistical Language and Speech Processing*, volume 10583, pages 94–105. Springer International Publishing, Cham.

Rozado, D. (2023). The Political Biases of ChatGPT. *Social Sciences*, 12(3):148.

Rubin, V., Conroy, N., Chen, Y., and Cornwell, S. (2016). Fake news or truth? using satirical cues to detect potentially misleading news. In *Proceedings of the Second Workshop on Computational Approaches to Deception Detection*, pages 7–17, San Diego, California. Association for Computational Linguistics.

Rüdiger, S. and Dayter, D. (2020). Manbragging online: Self-praise on pick-up artists' forums. *Journal of Pragmatics*, 161:16–27.

Ruggeri, F., Mesgar, M., and Gurevych, I. (2023). A Dataset of Argumentative Dialogues on Scientific Papers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7684–7699, Toronto, Canada. Association for Computational Linguistics.

Sagae, K. (2009). Analysis of discourse structure with syntactic dependencies and data-driven shift-reduce parsing. In *Proceedings of the 11th International Conference on Parsing Technologies (IWPT'09)*, pages 81–84.

Sapountzi, A. and Psannis, K. E. (2018). Social networking data analysis tools & challenges. *Future Generation Computer Systems*, 86:893–913.

Sardianos, C., Katakis, I. M., Petasis, G., and Karkaletsis, V. (2015). Argument Extraction from News. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 56–66, Denver, CO. Association for Computational Linguistics.

Schmidt, K., Niekler, A., Kantner, C., and Burghardt, M. (2023). Classifying Speech Acts in Political Communication: A Transformer-based Approach with Weak Supervision and Active

Learning. In *2023 18th Conference on Computer Science and Intelligence Systems (FedCSIS)*, pages 739–748.

Schneider, J. (2014). Automated argumentation mining to the rescue? Envisioning argumentation and decision-making support for debates in open online collaboration communities. In *Proceedings of the First Workshop on Argumentation Mining*, pages 59–63, Baltimore, Maryland. Association for Computational Linguistics.

Searle, J. R. (1969). *Speech acts: An essay in the philosophy of language*, volume 626. Cambridge university press.

Searle, J. R. (1976). A classification of illocutionary acts1. *Language in society*, 5(1):1–23.

Sellars, W. (1954). Some reflections on language games. *Philosophy of Science*, 21(3):204–228.

Sellars, W. (1969). Language as thought and as communication. *Philosophy and Phenomenological Research*, 29(4):506–527.

Shahsavari, S., Holur, P., Wang, T., Tangherlini, T. R., and Roychowdhury, V. (2020). Conspiracy in the time of corona: automatic detection of emerging covid-19 conspiracy theories in social media and the news. *Journal of computational social science*, 3(2):279–317.

Shi, B., Lam, W., Bing, L., and Xu, Y. (2016). Detecting common discussion topics across culture from news reader comments. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 676–685.

Shu, K., Wang, S., and Liu, H. (2019). Beyond news contents: The role of social context for fake news detection. In *Proceedings of the twelfth ACM international conference on web search and data mining*, pages 312–320.

Sia, S., Dalmia, A., and Mielke, S. J. (2020). Tired of topic models? clusters of pretrained word embeddings make for fast and good topics too! *arXiv preprint arXiv:2004.14914*.

Sia, S. and Duh, K. (2021). Adaptive mixed component lda for low resource topic modeling. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2451–2469.

Silva, A. (2016). Opinion Manipulation in Social Networks.

Sperrle, F., Sevastjanova, R., Kehlbeck, R., and El-Assady, M. (2019). VIANA: Visual Interactive Annotation of Argumentation.

Srivastava, A. and Sutton, C. (2017). Autoencoding variational inference for topic models. *arXiv preprint arXiv:1703.01488*.

Stede, M. and Neumann, A. (2014). Potsdam commentary corpus 2.0: Annotation for discourse research. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 925–929, Reykjavik, Iceland. European Language Resources Association (ELRA).

Stede, M., Taboada, M., and Das, D. (2017). Annotation guidelines for rhetorical structure. *Manuscript. University of Potsdam and Simon Fraser University*.

Stefanov, P., Darwish, K., Atanasov, A., and Nakov, P. (2020). Predicting the Topical Stance and Political Leaning of Media using Tweets. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 527–537, Online. Association for Computational Linguistics.

Taylor, S. (2013). *What is discourse analysis?* Bloomsbury Academic.

Tofiloski, M., Brooke, J., and Taboada, M. (2009). A syntactic and lexical-based discourse segmenter. In *Proceedings of the ACL-IJCNLP 2009 conference short papers*, pages 77–80.

Toledo-Ronen, O., Bar-Haim, R., and Slonim, N. (2016). Expert Stance Graphs for Computational Argumentation. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 119–123, Berlin, Germany. Association for Computational Linguistics.

Ulum, M., Sutopo, D., and Warsono, W. (2018). A comparison between trumpâ€™s and clintonâ€™s commissive speech act in americaâ€™s presidential campaign speech. *English Education Journal*, 8(2):221–228.

Uscinski, J. E. (2018). *Conspiracy theories and the people who believe them*. Oxford University Press, USA.

Ushio, T., Shi, H., Endo, M., Yamagami, K., and Horii, N. (2017). Recurrent convolutional neural networks for structured speech act tagging. *2016 IEEE Workshop on Spoken Language Technology, SLT 2016 - Proceedings*, pages 518–524.

van Eemeren, F. and van Haaften, T. (2023). The Making of Argumentation Theory: A Pragma-dialectical View. *Argumentation*, 37:1–36.

Vecchi, E. M., Falk, N., Jundi, I., and Lapesa, G. (2021). Towards Argument Mining for Social Good: A Survey. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1338–1352, Online. Association for Computational Linguistics.

Visser, J., Konat, B., Duthie, R., Koszowy, M., Budzynska, K., and Reed, C. (2020). Argumentation in the 2016 US presidential elections: Annotated corpora of television debates and social media reaction. *Language Resources and Evaluation*, 54(1):123–154.

Visser, J. and Lawrence, J. (2022). The skeptic web service: Utilising argument technologies for reason-checking. *Frontiers in Artificial Intelligence and Applications*, 353:375–376.

Visser, J., Lawrence, J., Reed, C., Wagemans, J., and Walton, D. (2021). Annotating Argument Schemes. *Argumentation*, 35.

Visser, J., Lawrence, J., Wagemans, J., and Reed, C. (2018). Revisiting computational models of argument schemes: 7th International Conference on Computational Models of Argument,

COMMA 2018. *Computational Models of Argument - Proceedings of COMMA 2018*, 305:313–324.

Vulić, I., De Smet, W., Tang, J., and Moens, M.-F. (2015). Probabilistic topic modeling in multilingual settings: An overview of its methodology and applications. *Information Processing & Management*, 51(1):111–147.

Wachsmuth, H., Naderi, N., Hou, Y., Bilu, Y., Prabhakaran, V., Thijm, T. A., Hirst, G., and Stein, B. (2017). Computational Argumentation Quality Assessment in Natural Language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 176–187, Valencia, Spain. Association for Computational Linguistics.

Wagemans, J. H. M. (2014). Argumentatie en debat. *Boom Lemma*.

Wagemans, J. H. M. (2023). How to identify an argument type? On the hermeneutics of persuasive discourse. *Journal of Pragmatics*, 203:117–129.

Walton, D. (2007). *Media Argumentation: Dialectic, Persuasion and Rhetoric*. Cambridge University Press.

Walton, D., Reed, C., and Macagno, F. (2008). *Argumentation Schemes*. Cambridge University Press.

Walton, D. and Reed, C. A. (2005). Argumentation Schemes and Enthymemes. *Synthese*, 145(3):339–370.

Wang, W. Y. (2017). " liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*.

Wang, Y., Kühn, R., Harris, R. A., Mitrovic, J., and Granitzer, M. (2022). Towards a unified multilingual ontology for rhetorical figures. In *KDIR*, pages 117–127.

Wen, S. (2016). A Pragma-Dialectical Approach to Political Discourse Analysis: A Case Study of the Former United State Trade Representative Ron Kirk's Remarks on the Poultry Case. *Sinología hispánica. China Studies Review*, 3(2):53–66.

Wilson, R. A. and Land, M. K. (2020/2021). Hate Speech on Social Media: Content Moderation in Context. *Connecticut Law Review*, 52:1029.

Wittgenstein, L. (2019). *Philosophical investigations*.

Wu, X., Dong, X., Nguyen, T., Liu, C., Pan, L.-M., and Luu, A. T. (2023). Infoctm: A mutual information maximization perspective of cross-lingual topic modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13763–13771.

Xia, M., Deng, W., Zhang, S., Liu, M., Xu, J., and Zhai, P. (2022). Automatic Recognition of Speech Acts in Classroom Interaction Based on Multi-Text Classification. *IEIR 2022 - IEEE International Conference on Intelligent Education and Intelligent Research*, pages 241–246.

Xie, Q., Zhang, X., Ding, Y., and Song, M. (2020). Monolingual and multilingual topic analysis using lda and bert embeddings. *Journal of Informetrics*, 14(3):101055.

Yue, Z., Zeng, H., Zhang, Y., Shang, L., and Wang, D. (2023). Metaadapt: Domain adaptive few-shot misinformation detection via meta learning. *arXiv preprint arXiv:2305.12692*.

Zeldes, A., Liu, Y. J., Iruskieta, M., Muller, P., Braud, C., and Badene, S. (2021). The disrpt 2021 shared task on elementary discourse unit segmentation, connective detection, and relation classification. In *Proceedings of the 2nd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2021)*, pages 1–12.

Zenker, F., van Laar, J. A., Cepollaro, B., Gâţă, A., Hinton, M., King, C. G., Larson, B., Lewiński, M., Lumer, C., Oswald, S., Pichlak, M., Scott, B. D., Urbański, M., and Wagemans, J. H. M. (2023). Norms of Public Argumentation and the Ideals of Correctness and Participation. *Argumentation*.

Zhang, R., Gao, D., and Li, W. (2011). What Are Tweeters Doing: Recognizing Speech Acts in Twitter.

Zhang, Z., Fang, M., Chen, L., and Namazi Rad, M. R. (2022). Is neural topic modelling better than clustering? an empirical study on clustering with contextual embeddings for topics. In Carpuat, M., de Marneffe, M.-C., and Meza Ruiz, I. V., editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3886–3893, Seattle, United States. Association for Computational Linguistics.

Zhao, H., Phung, D., Huynh, V., Jin, Y., Du, L., and Buntine, W. (2021). Topic modelling meets deep neural networks: a survey. In Zhou, Z.-H., editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, IJCAI International Joint Conference on Artificial Intelligence, pages 4713–4720, United States of America. Association for the Advancement of Artificial Intelligence (AAAI).

Zhao, W., Strube, M., and Eger, S. (2022). Discoscore: Evaluating text generation with bert and discourse coherence. *arXiv preprint arXiv:2201.11176*.

Zhu, Y., Sheng, Q., Cao, J., Li, S., Wang, D., and Zhuang, F. (2022). Generalizing to the future: Mitigating entity bias in fake news detection. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2120–2125.