Social Network Analysis and Mining (2024) 14:51
https://doi.org/10.1007/s13278-024-01212-1

ORIGINAL ARTICLE

Structural complexity predicts consensus readability in online discussions

Rafik Hadfi[1] · **Takayuki Ito**[1]

Received: 28 February 2023 / Revised: 29 December 2023 / Accepted: 26 January 2024

Abstract
The intricate relationship between structure and function spans various disciplines, from biology to management, offering insights into predicting interesting features of complex systems. This interplay is evident in online forums, where the organization of the threads interacts with the message's meaning. Assessing readability in these discussions is vital for ensuring information comprehension among diverse audiences. This assessment is challenging due to the complexity of natural language compounded by the social and temporal dynamics within social networks. One practical approach involves aggregating multiple readability metrics as a consensus alignment. In this study, we explore whether the structural complexity of online discussions can predict consensus readability without delving into the semantics of the messages. We propose a consensus readability metric derived from well-known readability tests and a complexity metric applied to the tree structures of Reddit discussions. Our findings indicate that this proposed metric effectively predicts consensus readability based on the complexity of discourse structure.

Keywords Social network analysis · Information and web mining · Structural complexity · Consensus alignment · Collective intelligence · Natural language processing · Discourse quality index

1 Introduction

Findings across disciplines ranging from biology to management consistently show a fundamental link between the structure of a system and its function (Carley 1995). Characterizing such structures could however be a challenging task for a number of reasons. The structure may not be fully defined and its constituents may be unknown to the observer. There exists a wide range of methodologies that could be used to quantify the structure of any given organization regardless of its constituents. The analysis becomes more complex when examining structures that exist at different scales and possess rich content.

Genomic data stands as a prime example for its intricate scales and variations among individuals and tissues (Morganti et al. 2019). Similarly, online forums such as Reddit or Quora exemplify complex socio-technical systems. Such platforms organize discussions into threads, enabling diverse viewpoints, nested replies, and the emergence of subtopics within conversations.

The study of online discussions often involves the qualitative analysis of the discourse by probing how textual content contributes to coherence and to the social function of the discourse (Johnstone 2017). Examining the readability of the text is one other way of ensuring that information is comprehensible across diverse audiences. Such approach could rely on quantitative measures, such as Flesch-Kincaid tests or Gunning fog index, which often evaluate text ease regardless of the underlying semantics (Beier et al. 2022). Such readability measures vary depending on the domain but could be applied in the form of consensus where a cohort of experts would naturally agree on the quality of a piece of text.

In online discussions, it is commonplace for textual messages to undergo multiple edits, thereby altering their temporal (and causal) sequence and complicating discourse and readability analysis. This challenge is especially notable when messages are deleted or vandalized (de Laat 2016).

To address this problem, we follow the intuition that, similar to various complex systems, the structure of an online discussion might offer insights into how readers collectively perceive its readability. We therefore propose a method to

✉ Rafik Hadfi
rafik.hadfi@i.kyoto-u.ac.jp

[1] Department of Social Informatics, Graduate School of Informatics, Kyoto University, Kyoto, Japan

Published online: 04 March 2024

predict the *consensus readability* of online discussions by looking at their structural complexity. This pursuit delves into the relationship between structure, consensus alignment, and ultimately, the collective intelligence of the authors (and readers) of the content.

We define readability as a consensus metric across well-known readability measures. We adopt such consensus approach to tackle the problem of quantifying content that could have various interpretations similar to how consensus is built in social settings. The idea of using consensus aggregation mechanisms is found in various domains ranging from politics (Van Gunten et al. 2016) to molecular biology (Schneider 2002). We then quantify the complexity of a discussion tree using information entropy applied to the structure of the tree (Shannon 1948). Our results show that the proposed complexity metric predicts the consensus readability of textual discussions on Reddit. This is the first study that combines the structural features of a discourse alongside its typed acts to produce a highly predictive model of the judgment that humans would attribute to it. Lexical, syntactic, and discourse factors have previously been used for similar task but do not account for the structure of discourse (Pitler and Nenkova 2008).

The article is structured as follows. In Sect. 2, we visit the related work on the quantification of complexity in various systems as well as various discourse analysis methodologies. In Sect. 3, we introduce the type of content we are targeting in the study. In Sect. 4, we introduce the formal method to represent and quantify structural complexity on discussions. Finally, we provide the experimental results and conclude.

2 Related work

There is an increasing number of theoretical and empirical studies showing a connection between the structure of a system and its function (Carley 1995). This is encountered for instance in material science (Callister Jr 2003), chemistry (Dickson 2011), biology (Honey et al. 2010; Bojar 2020; David 2003), architecture (Greenough 2020; Givoni 1998), organizational management (Chappell and Dewey 2014; San Cristóbal 2022), and linguistics (Van Valin Jr 2003).

Such connections are often unraveled using qualitative and quantitative tools from graph theory (Greenough 2020), physics (Fabac and Stepanić 2008), or information theory (Morzy et al. 2017; Schlick et al. 2013). One could for example rely on hierarchical measures for complex networks (Mones et al. 2012). In (Marin et al. 2022), the authors use the concept of mobility entropy and applied it to spatial interactions in urban city centers. Similarly, the authors in (Broniatowski and Moses 2014) looked at flexibility, complexity, and controllability in large scale systems. Another hierarchic metric in (Zamani et al. 2019) characterizes the differences in structure and the dynamics of networks retrieved from dark and public Web forums. Analyzing the content of Web forums extends beyond their structural aspects, offering valuable insights into understanding public discourse and its societal impact.

Discourse analysis is one way to analyze the content of online forums as it applies to monolithic blocks of text or threads of discussions (Johnstone 2017; Steenbergen et al. 2003). It generally relies on a variety of approaches, including critical discourse analysis, conversation analysis, ethnography, interactional sociolinguistics, and other qualitative or quantitative methods (Johnstone 2017).

Discourse analysts often look at how discourse segments contribute to the coherence of an overall content (Dontcheva-Navratilova and Povolná 2020; Rohde et al. 2018). They also look at the social, or deliberative, function of the content (Bächtiger and Parkinson 2019; Fournier-Tombs and Di Marzo Serugendo 2020; Hadfi and Ito 2022).

The deliberative approach goes beyond the linguistic aspects of discussions by looking at macroscopic factors that influence the evolution of the text. The authors in (Shin and Rask 2021) propose deliberative indicators based on a combination of networks and time-series analysis with the motivation of helping to monitor how online deliberations evolve. They adopted Habermasian deliberative criteria encompassing six throughput indicators, applying them to a participatory budgeting project in Finland (Habermas 2004). Similarly, the authors in (Steenbergen et al. 2003) propose a discourse quality index (DQI) as a quantitative measure of discourse in deliberation. The proposed index is also rooted in the discourse ethics of Habermas and gives an accurate representation of the most important criteria underlying deliberation.

There are various ways to aggregate DQI indicators with potential validity issues ranging from the omission of argumentative quality to insufficient sensitivity to context (Bächtiger et al. 2022; Bächtiger and Parkinson 2019). One way to circumvent such limitations is to apply quantitative approaches to measuring the discourse quality using machine learning techniques (Fournier-Tombs and Di Marzo Serugendo 2020).

The application of the previous tools often assume that the content is well-defined and unambiguous. It is however the case that discussion threads lack temporal structures due to concurrent modifications that may happen at various times after the content's creation. This makes conducting causal discourse analysis challenging. For instance, a user might edit their reply to another user, disrupting the textual connection between the messages and the temporal sequence. The situation becomes more complicated when a user, for instance, modifies their stance in an argument, resulting in an incomprehensible gap in the discourse. While these actions can be identified through persistent

discussion transcripts, they inevitably obscure understanding for other users during the discussions. Similarly, the spatial structure of the content could be altered through the deletion of some messages in the threads.

In the absence of clear spatiotemporal structure, it is possible to microscopically look at the properties of the content given stylistic or semantic aspects (Hadfi et al. 2022). The property we are interested in is that of readability. Monitoring the readability of the content on online forums is important because it ensures that information is comprehensible to a diverse audience, thereby promoting effective communication and facilitating the exchange of ideas. However, assessing such property is daunting because of the challenges that natural language poses, particularly when combined with the social and temporal dynamics of social networks. Being able to quantify readability independently from its linguistic substrate could remediate at these limitations.

Readability measures quantify the ease with which a reader can apprehend a written text. Examples of such measure include the Flesch-Kincaid readability tests (Kincaid et al. 1975), Gunning fog index (Powers et al. 1958), SMOG index (Hedman 2008), Coleman-Liau index (Coleman and Liau 1975), Automated Readability Index (Senter and Smith 1967), Linsear Write (McCannon 2019), and Dale-Chall readability (Stocker 1971). Such measures refer to various aspects to qualify written text, regardless of the semantics or discourse. They could for example describe the use of difficult words or the number of composed sentences. In some sense, they could be thought of as different experts asked to collectively decide on the readability of text. One question arises now, on the level of alignment between these different measures and whether some consensus could be reached, or not, when these "experts" are asked. This problem could be described as a *consensus alignment*, often used with genomic data (David 2003), but is extensible to the case of readability of textual content, namely *consensus readability*. We adopt such consensus approach to tackle the problem of quantifying the readability of a content that could have various interpretations similar to how consensus is built in social settings (Engel et al. 2014; Calof et al. 2022; Kabo 2018).

In the absence of a clear structure, aggregating consensus readability becomes challenging, particularly when no spatiotemporal connections exist between text blocks. Social network discussions often face issues like corrupted timestamps or vandalized content (de Laat 2016). This research seeks to determine if combining loose structural features at the macroscopic level with discourse types at the microscopic scale can predict the overall consensus readability. For instance, can we quickly predict the readability of a lengthy Reddit discussion as accurately as a group of experts would? Beyond the correlation of structure to function, this inquiry delves into how collective intelligence might manifest in specific structures or even communities (Heylighen 1999).

3 Discourse act discussions

3.1 Identifying discourse acts

The notion of discourse has been defined in numerous ways in linguistics. It is broadly referred to as discourse acts, or speech acts, when it occurs in spoken dialog (Johnstone 2017). Discourse acts describe how we meaningfully relate and categorize spoken natural language segments to achieve a certain performative function or action in communication. There exists many categories of discourse acts for conversation, such as Dialog Act Markup in Several Layers (DAMSL) (Core and Allen 1997; Stolcke et al. 1998), dialogAct Markup Language (DiAML) (Bunt et al. 2012) for spoken discourse, and Rhetorical Structure Theory (RST) for the argumentation within a single document (Mann and Thompson 1987). Compound discourses are also found in dialogs and consist of Narrative Discourse and Repartee Discourse (Larson 1984). Narrative discourse focuses on the depiction of motion and repartee discourse describes speech exchanges. The rise of social media has opened more and more room for people for the expression of opinions, ideas and arguments on diverse topics, and thus creating a new type of repartee discourse and opened new ways of understanding how people engage in discussions.

The way people engage in discussions is non homogenous across online media. While Facebook or Twitter are more used for the expression of opinions, platforms like Reddit or other community forums have a usage for querying and question answering. One way to understand the difference between these discussions is to study the high-level discourse structures. In these discourse structures, it is possible to assign tags called discourse acts tags to textual utterances with a particular function in the conversation. To illustrate this notion, let us take an examples of discourse acts in a discussion thread from the Reddit page "What is the most creative form of cheating you've seen?" shown in Fig. 1.

To obtain the discourse acts, we extract the elementary discourse units and then assign the underlying relationships. In dialogs, each utterance is connected to the utterance that it replies to. In our example of discussion in Fig. 1, the answer of Jane is linked to the question of Bob. Ron's answer is also linked to the question of Bob, and so on. In the case of structured platforms like Reddit, these relationships are already defined. Each comment on the platform is an explicit reply to another comment. But in Facebook or various group chat platforms, such structure is not explicitly defined. To decide which comment is directed in reply to whom, machine learning techniques will have to be adopted.

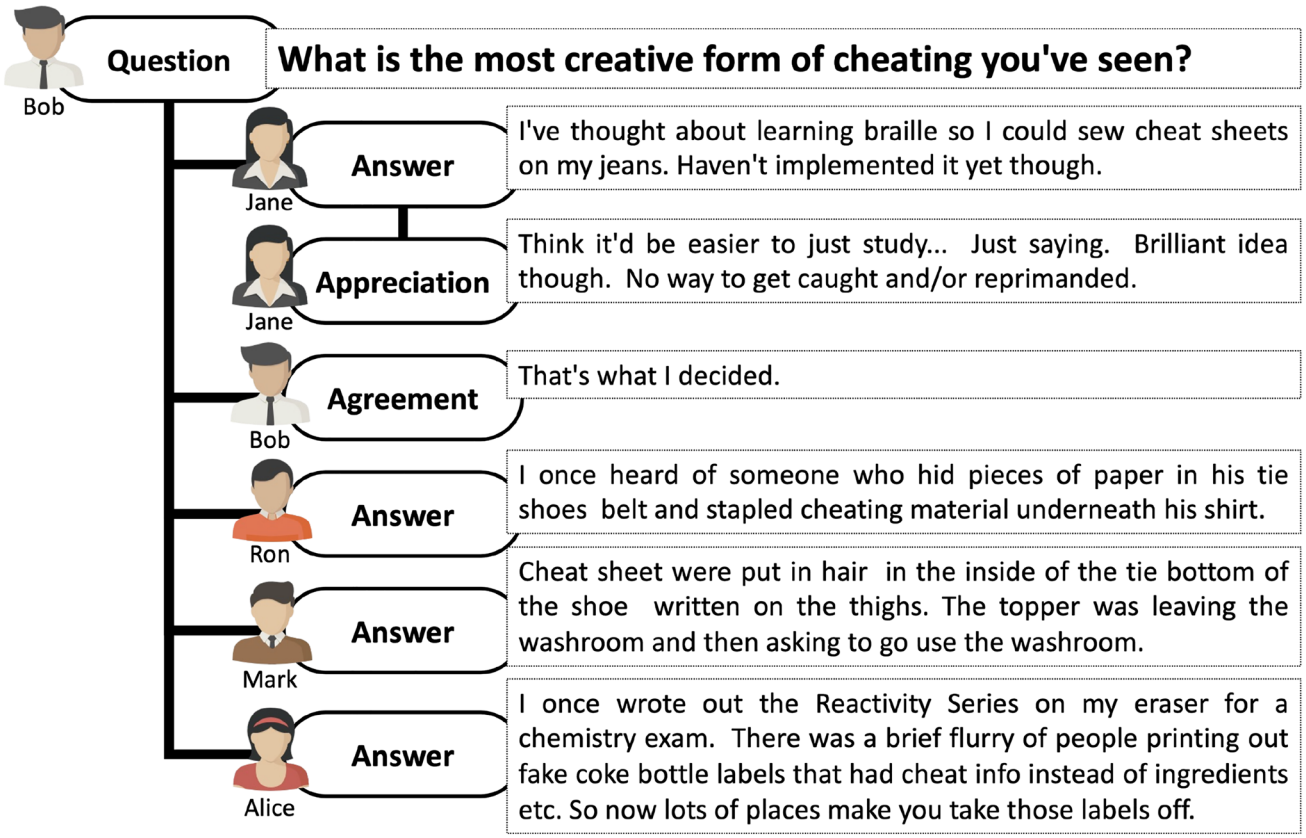

Fig. 1 Examples of discourse acts in a discussion thread from the Reddit page "What is the most creative form of cheating you've seen?"

For instance, (Dutta et al. 2019) adopted a Support Vector Machine approach and applied it to Facebook discussion threads. The approach we adopt herein for tagging the discourse acts of discussion comments relies on Conditional Random Fields (CRFs) (Zhang et al. 2017). The approach used a Reddit dataset with over 9K discussion threads with over 100K comments. Each comment is classified into one of ten different discourse act tags, namely, Announcement, Question, Answer, Elaboration, Humor, Agreement, Disagreement, Appreciation, Negative reaction, and undecidable categories as one tag. In the following, we adopt the same classification and dataset. In the example, once the comments are tagged, a discussion thread becomes a tree with the first or opening comment being the root node. A depth-first traversal of the tree gives multiple linear sequences of comments where each comment is posed as a reply to its previous one.

3.2 Categories of discourse acts

Quantifying the complexity of a discussion tree relies on its structure as well as the types of its nodes. The types will be defined as discourse act tags and will be assigned automatically. Discourse act tags play an important role in the automatic retrieval of textual content. Natural Language Processing (NLP) practitioners are interested in extracting such useful information and linking it to readability for example. This is particularly the case in open online communities such as Reddit (Tan et al. 2016). Such communities are not restricted for example to Q&A content (Cong et al. 2008; Hong and Davison 2009) or specific areas such as technical support sites (Kim et al. 2010). Here, we focus on wider range of online communities with the richness of the adopted discourse act tags. The discourse acts (Θ) that we will use when qualifying the complexity of discussion trees are summarized as follows (Zhang et al. 2017).

1. **Question** is a piece of information in the discourse that seeks some form of feedback. Questions do not necessarily require a question mark but could be posed in the form of a statement that is soliciting an answer. Moreover, rhetorical questions are not perceived as questions.
2. **Answer** responds to a particular question. It is possible to have multiple answers within the same discussion.
3. **Announcement** provides new information to the discussants, such as news, opinions, reviews, or general remarks.

4. **Agreement** is a comment that expresses agreement with some information provided in a previous comment. It can be an agreement with a point, a statement with supporting evidence, a positive example, or a confirmation or acknowledgment of a previously made point.
5. **Appreciation** is a comment that expresses excitement or praise in reply to another comment. Unlike agreements, appreciations do not evaluate the merits of the points being made.
6. **Disagreement** is a comment that aims at correcting, criticizing, contradicting, or objecting to a point. A disagreement can also provide evidence to support its motive such as an example or contrary anecdote.
7. **Negative reaction** is expressed to a previous comment by attacking it or mocking its author, or expressing emotions such as disgust, derision, or anger, to the contents of the targeted comment.
8. **Elaboration** adds additional information to the end of the comment it elaborates on. An author or a moderator might, for instance, elaborate on their question to provide more context, or elaborate on an answer to add additional information.
9. **Humor** is primarily a joke, a piece of sarcasm that is not necessarily trying to add information to the discussion. If a comment is sarcastic but uses sarcasm to make a point or provide information, then this comment may belong to a different category.
10. **Other** is used when encountering utterances that cannot be classified into the previous types. See also "undecided roles" (Dutta et al. 2019).
11. **Undefined** is used when encountering a content that cannot be treated as text. The "Undefined" tag generally accounts for non-textual or corrupt content. It is often common that users attach images or links along their textual messages. Such context is omitted during the parsing of the text.

Extending beyond the previous discourse act category depends on the reliability of our classification algorithm. Recent advancements in large-scale Language Models (LLMs) offer the possibility to define more intricate discourse tags, potentially incorporating complex semantic relations (Sun et al. 2023).

4 Methodology

4.1 Discourse act tree

The thread of an online discussion could be perceived as a discourse act tree (DAT) where each node is a message, or post, linked to other messages in a hierarchical manner. To quantify the complexity of a discussion tree, we need to find the number of possible microscopic states that can be assigned to the nodes of the tree. In our case, each node takes a value in the set of discourse types Θ, and could be an announcement, question, answer, elaboration, humor, agreement, disagreement, appreciation, or a negative reaction. Formally, a DAT is defined as follows.

Definition 1 (**DAT**) A Discourse Act Tree is a directed tree $\mathcal{T} = (V, E, \Theta)$, comprising a set of nodes V associated with discourse types Θ, and edge set E. Each node $v_{t,i} \in V(\mathcal{T})$ denotes a textual message written by user u at time t and having discourse type $\theta_{t,i} \in \Theta$.

The root node of the tree typically constitutes the initial post in the discussion, commonly an announcement or a question initiated by the moderator of the discussion. The edges within E illustrate the semantic structure of the tree, which might not always align with the actual thread progression in the discussion. This accounts for situations where a message replies to another message without being directly attached to it in the discussion tree.

The discourse types Θ are automatically estimated using classification techniques that assign a probability distribution $\mathbb{P}_\theta$ to each node (Hadfi et al. 2021; Ito et al. 2021). The discourse type of a textual message is not always guaranteed to be of a singular type. Hence, it should be defined probabilistically to accommodate various interpretations that a discourse analyst could attribute to the message within the context of the discussion.

The discussion tree $\mathcal{T}$ has L levels starting from the leaf posts at the bottom ($\ell = 1$) up to the the initially posted root node ($\ell = L$). Here, we assume that there are M posts in total, distributed across the L levels. Note that the set $\mathcal{T}_\ell$ is the set of nodes located on level ℓ. The cardinality of $\mathcal{T}_\ell$ is the number of posts at level ℓ, given by $M^{(\ell)}$, so that Eq. (1) holds.

$$M = \sum_{\ell=1}^{L} M^{(\ell)} \tag{1}$$

The discussion tree is organized in such a manner that each post at the ℓth level is connected to its parent post at the $\ell + 1$ level. There are $M_n^{(\ell)}$ posts located at level ℓ and connected to their parent node n located at level $\ell + 1$. Summing up all posts for layer ℓ gives Eq. (2).

$$M^{(\ell)} = \sum_{n \in \mathcal{T}_{\ell+1}} M_n^{(\ell)} \tag{2}$$

Equations (1) and (2) encapsulate the inter-level connectivity in a DAT and will be used in the next section to compute the number of micro-states of the tree.

4.2 Complexity of a discourse act tree

To quantify the complexity of a discussion tree $\mathcal{T}$, we need to find the number of possible ways to organize the M posts in the tree across the L layers, and while assuming that each node could take some discourse type from the set Θ. This is a combinatorial problem and the number of possible combinations, $\Omega_{\mathcal{T}}$, is computed using the multinomial form in Eq. (3).

$$\Omega_{\mathcal{T}} = \frac{M!}{\prod_{\ell=1}^{L} M^{(\ell)}!} \prod_{\ell=1}^{L} \frac{M^{(\ell)}!}{\prod_{n \in \mathcal{T}_{\ell+1}} \prod_{\theta \in \Theta} M_{n,\theta}^{(\ell)}!} \tag{3}$$

Here, $M_{n,\theta}^{(\ell)}$ is the number of posts of type θ located at level ℓ and connected to their parent node n located at level $\ell + 1$. The components of the factorials and products in Eq. (3) capture the combinations and permutations of these arrangements across the layers and nodes, determining the total number of possible ways to organize the variables within $\mathcal{T}$.

To quantify the complexity of an online discussion, we propose to look at the number of states that all of the connected nodes of the discussion tree could take. To this end, we will borrow a concept from statistical mechanics that specifies the relationship between entropy and the number of possible micro-states of a system (Perrot 1998). The entropy S is proportional to the natural logarithm of the number of micro-states, Ω, illustrated in Eq. (4).

$$S = k_B \log \Omega \tag{4}$$

where k_B is known as the Boltzmann constant (Perrot 1998). For instance, consider the discussion tree depicted in Fig. 1, comprising 6 replies to Bob's initial question. Each reply within this tree can be categorized into a type from the set Θ containing a total of 11 distinct types. With a total number of possible arrangements equal to $\Omega = 11^6$, and assuming a value of $k_B = 1$, the resulting entropy is $S = 6.24$.

The formulation of the structural complexity of any discussion tree $\mathcal{T}$ could be defined in the same way once applied to the number of possible micro-states $\Omega_{\mathcal{T}}$ that the tree nodes could take. That is, after substituting the logarithm of the factorials using a Stirling approximation (Robbins 1955), we obtain the analytical entropy of $\Omega_{\mathcal{T}}$, illustrated in Eq. (5),

$$\begin{aligned} S_{\mathcal{T}} &= \log \Omega_{\mathcal{T}} \\ &= \sum_{\ell=1}^{L} M^{(\ell)} \log P^{(\ell)} - \sum_{\ell=1}^{L} \sum_{n \in \mathcal{T}_{\ell+1}} \sum_{\theta \in \Theta} M_{n,\theta}^{(\ell)} \log P_{n,\theta}^{(\ell)} \end{aligned} \tag{5}$$

with $P^{(\ell)} = \frac{M^{(\ell)}}{M}$ and $P_{n,\theta}^{(\ell)} = \frac{M_{n,\theta}^{(\ell)}}{M^{(\ell)}}$. The term $M_{n,\theta}^{(\ell)}$ is the number of posts of type θ located at level ℓ and connected to their parent node n at level $\ell + 1$.

The first term in (5) could be interpreted as the vertical entropy (S_V) of the tree and the second term as the typed horizontal entropy ($S_{H\Theta}$). The two terms of the entropy describe how complexity arises across the two dimensions of the tree. The horizontal entropy is due to changes within the same level of the discussion and often reflect different takes on the same subject or content written in the parent post. On the other hand, the vertical entropy is due to changes between different levels of discussion and often occur at different time frames. For instance, the leaves of the discussion tree could deviate from earlier posts as they go into tangent topics of discussion.

4.3 Consensus readability

To assess the quality of any given text, we do not take one particular metric nor average across distinct metrics. Instead, we seek to build a consensus among metrics by picking the most common scores for a given text input. This procedure is used for identifying and resolving disagreements among different ways to interpret the readability of the same content. Consensus alignment mechanisms are found for instance in molecular biology (Schneider 2002). Herein, we consider a family of readability metrics, namely $\mathcal{M}$, often used in text and discourse analysis. This family is constituted of the Flesch-Kincaid readability test (Kincaid et al. 1975), Gunning fog index (Powers et al. 1958), SMOG index (Hedman 2008), Coleman-Liau index (Coleman and Liau 1975), Automated Readability Index (Senter and Smith 1967), Linsear Write (McCannon 2019), and Dale-Chall readability (Stocker 1971). The factors that differentiates these metrics are described in Table 1 ccording to how much they rely on content and/or shape.

The factors contributing to the construction of these metrics predominantly rely on quantifiable attributes within the text, such as the average length of sentences and words. To forge a unified metric surpassing individual measurements, we introduce the consensus readability, CR, as the subset of metrics from an original metric set $\mathcal{M}$ that align for a specific textual input taken from a node $v \in V(\mathcal{T})$.

Table 1 Taxonomy of the adopted readability metrics ($\mathcal{M}$)

Metrics	Description
Flesch-Kincaid readability test	Used words, sentences, and syllables
Gunning fog index	Words, sentences, and complex words
SMOG index	Number of sentences and polysyllables
Coleman-Liau index	Average number of letters and sentences per 100 words
Automated Readability Index	Characters, words, and sentences
Linsear Write	Easy words, hard words, and sentences
Dale-Chall readability	Words, difficult words, and sentences

The Eq. (6) calculates the consensus readability CR by essentially determining the set of readability metrics that yield the most frequent readability score.

$$CR(v) = \arg\max_{\mu \in \mathcal{M}} \sum_{\mu' \in \mathcal{M}} w_\mu \delta\big(\mu(v) - \mu'(v)\big) \tag{6}$$

with $\delta(\mu(v) - \mu'(v))$ being the Kronecker delta function that equals 1 if $\mu(v) = \mu'(v)$ and 0 otherwise. We also assign a weight $w_\mu \in [0, 1]$ to each metric $\mu \in \mathcal{M}$ to emphasize or de-emphasize its influence on the final consensus readability. Similar formulations are found in weighted majority voting or Ensemble Learning (EL) techniques in machine learning (Dogan and Birant 2019). For simplicity, we assume that $w_\mu = 1 \; \forall \mu \in \mathcal{M}$.

Now, for the full discourse act tree $\mathcal{T}$ and its combined textual content $V(\mathcal{T})$, the final readability measure $R(\mathcal{T})$ is defined in Eq. (7),

$$R(\mathcal{T}) = \sum_{v \in \mathcal{V}(\mathcal{T})} CR(v) \tag{7}$$

Table 2 Distribution of discourse act types in reddit discussions

Class	Number of instances
Announcement	1933
Question	16852
Answer	39734
Elaboration	18511
Disagreement	3284
Agreement	4871
Appreciation	8515
Negative reaction	1835
Humor	2352

where we apply the consensus readability metrics CR to the nodes of $V(\mathcal{T})$. In the following, we investigate how $R(\mathcal{T})$ relates to the structural complexity of a discourse act tree.

5 Results

In this section, we illustrate the predictiveness of our complexity measures give the consensus readability. To this end, we will use a discourse act discussion dataset extracted from Reddit and containing over 9K discussion threads with over 100K comments (Zhang et al. 2017). The original dataset used 9 discourse act tags and will define our types Θ. The tags are illustrated in Table 2.

We now look at how the structural entropies of any given discussion maps to the consensus readability of its content. After computing the consensus readability measures and the structural entropies, we look at the Pearson correlations between different entropies and the consensus readability measures as shown in Fig. 2.

After adding the θ term to the structural complexity, we obtain the correlation in Fig. 3. We observe a weaker positive correlation after integrating θ with the vertical entropy.

The correlation between the consensus readability and the entropies, shown in Table 3, indicate how predictable is the consensus readability given the structure of the tree and its entropic complexity. Balancing the readability consensus locally on the level of the discussion nodes and the global complexity of the tree could account for this correlation where distinct evaluators of readability reach a consensus that is also reflected in their diversity (S_{HU}) as well as in the diversity of the types of the nodes (S_θ). Herein, diversity is captured using the information theoretic notion of entropy (Marin et al. 2022; Morzy et al. 2017; San Cristóbal 2022).

The mechanisms that link structural complexity to consensus readability are mainly due to the cognitive roots of readability and how it relates to structure perception.

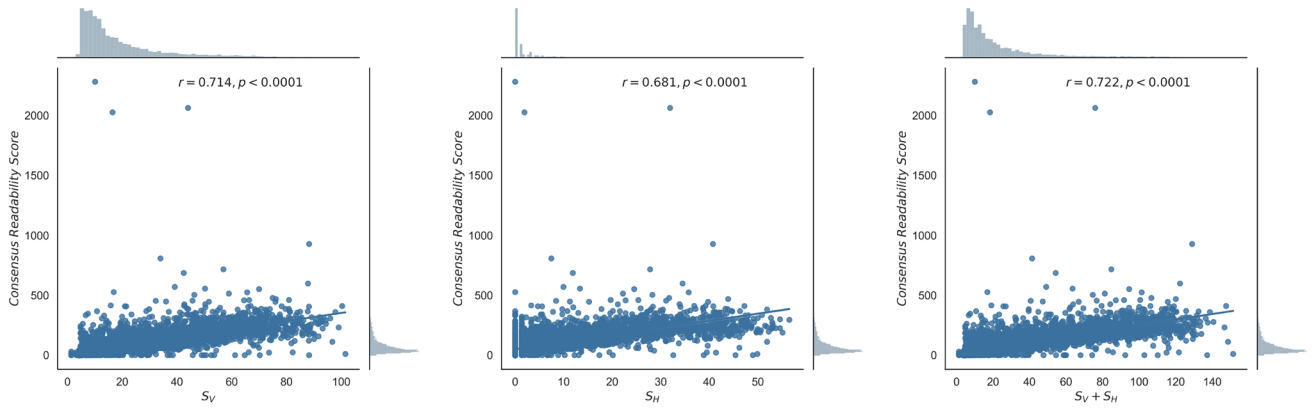

Fig. 2 Correlations between readability and non-typed complexity

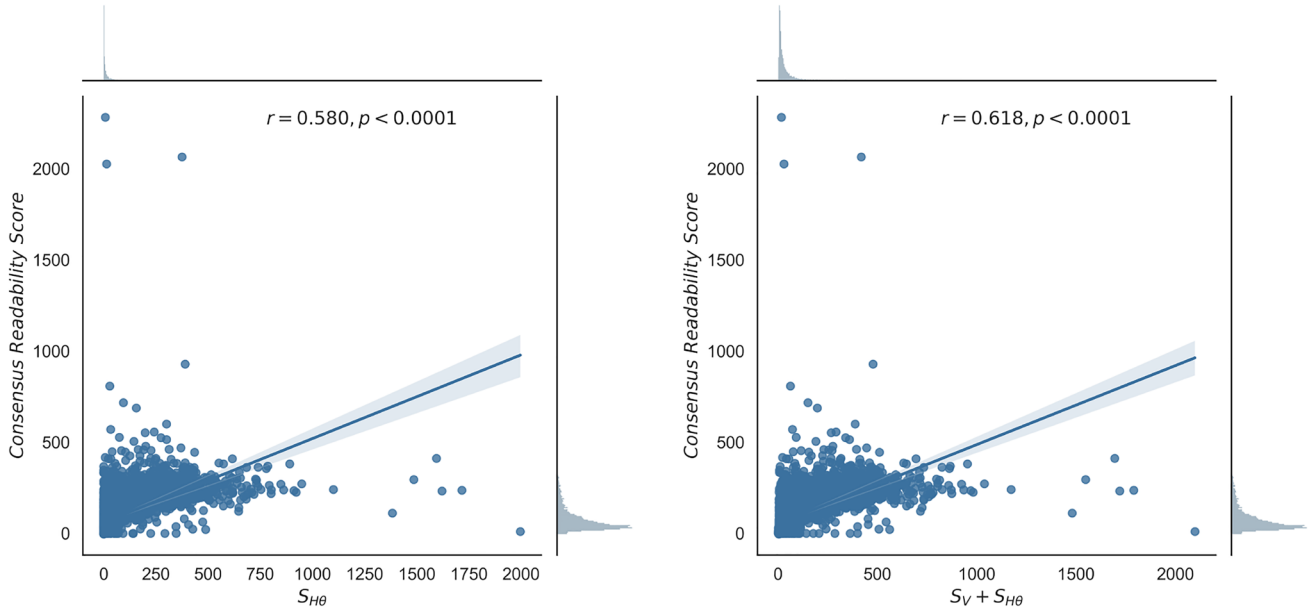

Fig. 3 Correlation between readability and typed complexity

Table 3 Correlation between readability and entropies ($p < .0001$)

Metric Type	Pearson Correlation
Vertical Entropy (S_V)	0.71
Horizontal Entropy (S_H)	0.68
Horizontal Entropy ($S_V + S_H$)	0.72
Typed Entropy ($S_{H\theta}$)	0.58
Full Entropy ($S_V + S_{H\theta}$)	0.61
Horizontal with users (S_{HU})	0.73

Readability is mainly influenced by factors that include the content (e.g., complexity of vocabulary) as well as the way this content is presented to the reader (e.g., font size, or spacing). These factors require different cognitive skills that affect visibility, speed of perception, fatigue in reading, or eye movements (Beier et al. 2022). These skills are indirectly captured by how readability metrics are designed (Beier et al. 2022; Pitler and Nenkova 2008) and how much they rely on content or shape as illustrated in Table 1. Since readability assesses the ease with which humans read and understand written texts, a structurally complex content that is linked by particular discourse acts becomes more difficult to process and comprehend. This is visible for instance in student online discussions (Polo and Varela 2018).

Finally, the proposed consensus readability (*CR*) in Eq. (6) is applied to one node at a time, independently from other nodes of the discussion tree. Readability assessments should not overlook how surrounding text coherence influences comprehension (Klare 1974). Recent advancements in natural language processing, particularly through the Transformer architecture, emphasize the role of context in improving readability prediction models (Meng et al. 2020). Introducing conditional dependence within *CR* can enhance its accuracy and relevance by considering the contextual relationships among nodes within the text structure. In this case, we will redefine the consensus readability as $CR(v \mid c_{\mathcal{T}}(v))$ where $c_{\mathcal{T}}(v)$ represents the context of node $v \in V(\mathcal{T})$. This conditional dependence on the context can be established in various ways. One could for example assess the readability of a node, considering the readability of its parent node(s) and how it could provide insights into the writing style of the author(s). Moreover, nodes within the same topic often share similarities in readability. For instance, within a paragraph discussing a specific theme (e.g., Law, or Mathematics), nodes might exhibit similar readability traits (e.g., legal terms, or formulas). Most importantly, messages written by the same author(s) might exhibit consistent readability patterns. Conditioning *CR* on the user that authored a given node could look at a her typical style when evaluating the consensus readability.

6 Conclusions

This study looks at the possibility of predicting the consensus readability of online discourse by looking primarily at its structure. We empirically demonstrate that discourse structure is strongly associated with its perceived readability. To this end, we quantified the structural complexity of the

discussion trees using information entropy. We then looked at readability as a consensus alignment that aggregates the readability of the content with multiple well-known readability tests (Kincaid et al. 1975; Powers et al. 1958; Hedman 2008; Coleman and Liau 1975; Senter and Smith 1967; McCannon 2019; Stocker 1971). The complexity metric is then applied to the discussion structure while accounting for the discussion discourse types. The complexity measure was tested on Reddit discussions and is shown to predict the readability of the Reddit content regardless of its underlying semantics.

One way to extend this work is to look at the context-dependent *CR* as pointed out in the previous section. Moreover, it is possible to investigate the temporal evolution of the entropy of DATs and whether it is predictive of the readability metric or not. Finally, we plan on testing our approach on more structured content such as Wikipedia (Ruprechter et al. 2020). Wikipedia articles possess a quality index that could potentially be studied using our structural metrics.

Author Contributions RH and TI ideated the research. RH performed the experiments and wrote the manuscript. TI provided feedback and editing suggestions. All authors approved the final version.

Funding This research was partially supported by JSPS Kakenhi Grant Number JP20K11936 and JST CREST Grant Number JPMJCR20D1.

Data availability The used dataset is publicly available (Zhang et al. 2017).

Code availability The code used to generate the results reported in the manuscript is available upon request.

Declarations

Conflict of interest The authors have no competing interests to declare on the content of this article.

References

Bächtiger A, Parkinson J (2019) Mapping and measuring deliberation: toward a new deliberative quality. Oxford University Press, Oxford

Beier S, Berlow S, Boucaud E et al (2022) Readability research: an interdisciplinary approach. Found Trends Human Comput Interact 16(4):214–324

Bojar D (2020) Structure determines function-the role of topology in the functionality of gene circuits. Synth Biol 5(1):ysaa008

Broniatowski DA, Moses J (2014) Flexibility, complexity, and controllability in large scale systems. Engineering systems division (ESD) Working Paper Series

Bunt H, Alexandersson J, Choe J, et al (2012) Iso 246170-2: A semantically-based standard for dialogannotation. In: proceedings of the 8th international conference on language resources and evaluation, Istanbul, Turkey, ELRA, p 8

Bächtiger A, Gerber M, Fournier-Tombs E (2022) 83Discourse Quality Index. In: research methods in deliberative democracy. Oxford University Press, https://doi.org/10.1093/oso/9780192848925.003.0006

Callister WD Jr (2003) Recovery, recrystallization, and grain growth. Materials science and engineering, an introduction,. Wiley, New Jersey, pp 180–184

Calof J, Søilen KS, Klavans R et al (2022) Understanding the structure, characteristics, and future of collective intelligence using local and global bibliometric analyzes. Technol Forecast Social Change 178(121):561

Carley KM (1995) Computational and mathematical organization theory: Perspective and directions. Computational & mathematical organization theory 1(1):39–56

Chappell D, Dewey TG (2014) Defining the entropy of hierarchical organizations. Compl Govern Netw 1(2):41–56

Coleman M, Liau TL (1975) A computer readability formula designed for machine scoring. J Appl Psychol 60(2):283

Cong G, Wang L, Lin CY et al (2008) Finding question-answer pairs from online forums. In: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, pp 467–474

Core MG, Allen J (1997) Coding dialogs with the damsl annotation scheme. In: AAAI fall symposium on communicative action in humans and machines, Boston, MA, pp 28–35

David W (2003) Mount. 2004. bioinformatics: Sequence and genome analysis. Gold Spring Harbor Laboratory press, New York pp 1–18

de Laat PB (2016) Profiling vandalism in wikipedia: a schauerian approach to justification. Ethics Inform Technol 18:131–148

Dickson RM (2011) Structure determines function in nanoparticles, their interfaces, and their assemblies

Dogan A, Birant D (2019) A weighted majority voting ensemble approach for classification. In: 2019 4th international conference on computer science and engineering (UBMK), IEEE, pp 1–6

Dontcheva-Navratilova O, Povolná R (2020) Coherence and cohesion in spoken and written discourse. Cambridge Scholars Publishing

Dutta S, Chakraborty T, Das D (2019) How did the discussion go: Discourse act classification in social media conversations. In: linking and mining heterogeneous and multi-view data. Springer: London p 137–160

Engel D, Woolley AW, Jing LX et al (2014) Reading the mind in the eyes or reading between the lines? theory of mind predicts collective intelligence equally well online and face-to-face. PloS one 9(12):e115,212

Fabac R, Stepanić J (2008) Modeling organizational design-applying a formalism model from theoretical physics. J Inform Organiz Sci 32(1):25–32

Fournier-Tombs E, Di Marzo Serugendo G (2020) Delibanalysis: understanding the quality of online political discourse with machine learning. J Inform Sci 46(6):810–822

Givoni B (1998) Climate considerations in building and urban design. Wiley, New Jersey

Greenough H (2020) Form and function. In: Form and Function. University of California Press
Habermas J (2004) Discourse ethics. In: Ethics: Contemporary Readings. Routledge, p 146–153
Hadfi R, Ito T (2022) Augmented democratic deliberation: Can conversational agents boost deliberation in social media? In: proceedings of the 21st international conference on autonomous agents and multiagent systems, pp 1794–1798
Hadfi R, Haqbeen J, Sahab S et al (2021) Argumentative conversational agents for online discussions. J Syst Sci Syst Eng 30:1–15
Hadfi R, Moustafa A, Yoshino K et al (2022) Best-answer prediction in q &a sites using user information. https://doi.org/10.48550/ARXIV.2212.08475
Hedman AS (2008) Using the smog formula to revise a health-related document. Am J Health Edu 39(1):61–64
Heylighen F (1999) Collective intelligence and its implementation on the web: algorithms to develop a collective mental map. Computat Math Organiz Theory 5:253–280
Honey CJ, Thivierge JP, Sporns O (2010) Can structure predict function in the human brain? Neuroimage 52(3):766–776
Hong L, Davison BD (2009) A classification-based approach to question answering in discussion boards. In: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, pp 171–178
Ito T, Hadfi R, Suzuki S (2021) An agent that facilitates crowd discussion. Group Decision and Negotiation pp 1–27
Johnstone B (2017) Discourse analysis. Wiley, New Jersey
Kabo F (2018) The architecture of network collective intelligence: correlations between social network structure, spatial layout and prestige outcomes in an office. Philosoph Trans Royal Soc B Biolog Sci 373(1753):20170,238
Kim SN, Wang L, Baldwin T (2010) Tagging and linking web forum posts. In: proceedings of the fourteenth conference on computational natural language learning, pp 192–202
Kincaid JP, Fishburne RP Jr, Rogers RL et al (1975) Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Tech. rep, Naval technical training command millington tn research branch
Klare GR (1974) Assessing readability. Read Res Quart 1:62–102
Larson ML (1984) Meaning based translation. University Press of America Lanham, MD
Mann WC, Thompson SA (1987) Rhetorical structure theory: a theory of text organization. University of Southern California, Information Sciences Institute Los Angeles
Marin V, Molinero C, Arcaute E (2022) Uncovering structural diversity in commuting networks: global and local entropy. Sci Rep 12(1):1–13
McCannon BC (2019) Readability and research impact. Econom Lett 180:76–79
Meng C, Chen M, Mao J et al (2020) Readnet: A hierarchical transformer framework for web article readability analysis. In: advances in information retrieval: 42nd European conference on IR research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part I 42, Springer, pp 33–49
Mones E, Vicsek L, Vicsek T (2012) Hierarchy measure for complex networks. PloS one 7(3):e33,799
Morganti S, Tarantino P, Ferraro E et al (2019) Complexity of genome sequencing and reporting: next generation sequencing (ngs) technologies and implementation of precision medicine in real life. Crit Rev Oncol Hematol 133:171–182
Morzy M, Kajdanowicz T, Kazienko P (2017) On measuring the complexity of networks: kolmogorov complexity versus entropy. Complexity 2017
Perrot P (1998) A to Z of thermodynamics. supplementary series; 27, Oxford University Press, URL https://books.google.co.jp/books?id=EBSbdNLmD-oC
Pitler E, Nenkova A (2008) Revisiting readability: a unified framework for predicting text quality. In: proceedings of the 2008 conference on empirical methods in natural language processing, pp 186–195
Polo FJF, Varela MC (2018) A structural analysis of student online forum discussions. In: languages at the crossroads: training, accreditation and context of use, Universidad de Jaén, pp 189–200
Powers RD, Sumner WA, Kearl BE (1958) A recalculation of four adult readability formulas. J Edu Psychol 49(2):99
Robbins H (1955) A remark on stirling's formula. Am Math Monthly 62(1):26–29
Rohde H, Johnson A, Schneider N et al (2018) Discourse coherence: Concurrent explicit and implicit relations. In: proceedings of the 56th annual meeting of the association for computational linguistics (Volume 1: Long Papers). Association for computational linguistics, Melbourne, Australia, pp 2257–2267, https://doi.org/10.18653/v1/P18-1210
Ruprechter T, Santos T, Helic D (2020) Relating wikipedia article quality to edit behavior and link structure. Appl Netw Sci 5(1):1–20
San Cristóbal J (2022) The network entropy as a measure of a complexity for project organizational structures. Proc Comput Sci 196:756–762
Schlick CM, Duckwitz S, Schneider S (2013) Project dynamics and emergent complexity. Computat Math Organiz Theory 19(4):480–515
Schneider TD (2002) Consensus sequence zen. Appl Bioinform 1(3):111
Senter R, Smith EA (1967) Automated readability index. Cincinnati Univ OH, Tech. rep
Shannon CE (1948) A mathematical theory of communication. Bell Syst Techn J 27(3):379–423
Shin B, Rask M (2021) Assessment of online deliberative quality: new indicators using network analysis and time-series analysis. Sustainability 13(3):1187
Steenbergen MR, Bächtiger A, Spörndli M et al (2003) Measuring political deliberation: a discourse quality index. Comparat Europ Polit 1(1):21–48
Stocker LP (1971) Increasing the precision of the dale-chall readability formula. Read Improve 8(3):87
Stolcke A, Shriberg E, Bates R et al (1998) Dialog act modeling for conversational speech. In: AAAI spring symposium on applying machine learning to discourse processing, pp 98–105
Sun X, Li X, Li J et al (2023) Text classification via large language models. arXiv:2305.08377
Tan C, Niculae V, Danescu-Niculescu-Mizil C et al (2016) Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In: proceedings of the 25th international conference on world wide web, pp 613–624
Van Gunten TS, Martin JL, Teplitskiy M (2016) Consensus, polarization, and alignment in the economics profession. Sociol Sci 3:1028–1052
Van Valin Jr RD (2003) Functional linguistics. The handbook of linguistics pp 319–336
Zamani M, Rabbani F, Horicsányi A et al (2019) Differences in structure and dynamics of networks retrieved from dark and public web forums. Phys A Statist Mech Appl 525:326–336
Zhang A, Culbertson B, Paritosh P (2017) Characterizing online discussion using coarse discourse sequences. In: proceedings of the international AAAI conference on web and social media