LLMs as Tools for Evaluating Textual Coherence: A Comparative Analysis

Bryan K. S. Barbosa[1], **Claudio E. C. Campelo**[1]

[1]Systems and Computing Department
Federal University Campina Grande (UFCG) – Campina Grande, PB – Brazil

bryankhelven@ieee.org, campelo@dsc.ufcg.edu.br

Abstract. *This study evaluate recent Large Language Models (LLMs), such as GPT-4o, GPT-3.5, Claude Opus, and LLaMA 2, for their ability to analyze textual coherence. The research focuses on three areas: local coherence, where models like GPT-4o and Claude Opus excel; global coherence, where Claude Opus is most effective; and incoherence detection, where GPT-4o demonstrates strong performance. These findings reveal both the capabilities and areas for improvement in current models, shedding light on their potential applications in natural language processing, paving the way for improvements in the field.*

Resumo. *Este estudo avalia o desempenho de Grandes Modelos de Língua (LLMs) recentes, como GPT-4o, GPT-3.5, Claude Opus e LLaMA 2, na análise automática de coerência textual. A pesquisa foca em três aspectos: coerência local, onde GPT-4o e o Claude Opus se destacam; coerência global, na qual Claude Opus é o mais eficaz; e detecção de incoerências, onde GPT-4o apresenta melhor desempenho. Esses resultados revelam as capacidades e limitações dos modelos atuais, contribuindo para o entendimento de suas aplicações no âmbito do Processamento de Línguas Naturais e trazendo avanços contínuos à área.*

1. Introduction

The concept of coherence lies at the very heart of effective communication, serving as a keystone element that determines the clarity, understandability, and overall quality of textual content [Koch and Travaglia 2003]. Coherence transcends the boundaries of syntax or grammar; it embodies the logical flow of ideas, ensuring that a text is not just a collection of sentences but a unified whole that conveys meaning with precision and subtlety [Freitas 2013]. As we move deeper into the digital age, where written text interactions are increasingly prevalent [Hoey 2013], the capacity to automatically analyze textual coherence became a crucial task within the realm of Natural Language Processing (NLP).

The advent of Large Language Models (LLMs) such as GPT-3, Llama, and Gemini has revolutionized our approach to generating text that mirrors the nuance and depth of human-written content. These models, trained on extensive corpora, have demonstrated an impressive ability to produce coherent and contextually relevant text across a wide range of topics. This proficiency in text generation naturally extends to the potential for these models to excel in tasks related to textual analysis. The underlying hypothesis is simple yet profound: if an LLM can generate coherent text, it should, by extension, possess a refined ability to discern coherence – or the lack thereof – in existing texts.

In the field of computational linguistics, textual coherence is defined by the logical and orderly sequence in which ideas are presented within a text, ensuring that information and arguments are conveyed in a comprehensible and fluid manner [Seno and Rino 2005]. This involves not only the superficial connection between sentences through discourse markers or transition words but also a deeper harmony in terms of theme, purpose, and shared knowledge between the author and the reader [Charolles 1978]. For NLP systems, assessing the coherence of a text implies understanding how its constituent parts – whether at the sentence, paragraph, or document level – come together to form a unified whole that is logically consistent and aesthetically pleasing [Jurafsky and Martin 2024]. This definition highlights the complexity of the textual coherence analysis task, underscoring it as a significant challenge within the field.

Historically, coherence has been conceptualized through various theoretical frameworks. Rhetorical Structure Theory (RST) [Mann and Thompson 1987] posits that text coherence is derived from the hierarchical organization of text units, while Centering Theory [Grosz et al. 1995] emphasizes the role of discourse entities and their continuity across sentences. Over time, computational approaches to coherence have evolved from rule-based systems, relying on explicit coherence markers and structural patterns, to sophisticated machine learning algorithms that infer coherence implicitly from large datasets [Jurafsky and Martin 2024]. The development of neural network-based models, particularly those employing attention mechanisms such as BERT [Devlin et al. 2018], has marked a significant advancement, enabling a deeper understanding of contextual relationships within texts.

Given this context, the primary objective of this study is to evaluate the capabilities of various LLMs in the analysis of textual coherence. Specifically, the study assesses how the models GPT 3.5, GPT 4, GPT 4o, Claude 3 Opus, Claude 3.5 Sonnet, Claude 3 Haiku, Gemini, LLaMA 2 13b, LLaMA 2 7b, and Bard perform in three key tasks: classifying texts as (i) locally or (ii) globally coherent or incoherent, and (iii) identifying specific incoherent segments within texts. By examining these aspects, the study aims to contribute to the ongoing dialogue on improving NLP technologies and advancing our understanding of how machines process and understand the subtleties of human language.

The remainder of this article is structured as follows: Section 2 reviews key theories and models in textual coherence analysis. Section 3 presents the relevant literature, while Section 4 details the methodology, including the models evaluated and the metrics used. Section 5 discusses the results and their implications, and Section 6 concludes with a summary of findings and suggestions for future research.

2. Theoretical Background

Textual cohesion and coherence are fundamental to discourse analysis and NLP, as they explain how texts are structured and interpreted. Cohesion refers to the connections within a text created through various linguistic relations, such as pronouns, conjunctions, and lexical ties, ensuring that the text is perceived as a unified whole rather than a random collection of sentences [Halliday and Hasan 1976]. Coherence, on the other hand, is a more abstract concept, referring to the logical and meaningful organization of ideas within a text, allowing readers to follow the flow of information and understand the intended message [Van Dijk 1977].

Cohesion can be achieved through grammatical and lexical means. Grammatical cohesion includes the use of pronouns, ellipses, and conjunctions to link sentences, while lexical cohesion involves the repetition of words or the use of synonyms to maintain the continuity of ideas. However, a text can be cohesive without being coherent if the sentences do not contribute to a meaningful whole [Koch and Travaglia 2003].

Coherence can be examined at two levels: local and global. Local coherence refers to the logical connections between adjacent sentences and paragraphs, ensuring that each idea flows smoothly into the next. This is often achieved through cohesive devices, such as pronouns and conjunctions, which help maintain continuity in meaning. Global coherence, on the other hand, concerns the overall structure and unity of the text, where all parts contribute to a consistent and meaningful whole [Charolles 1978]. Both levels of coherence are essential for a text to be understood as a cohesive and logically organized entity.

Theoretical frameworks like RST and Centering Theory have been foundational in the study of coherence. For instance, RST [Mann and Thompson 1987] analyzes the hierarchical organization of text by examining the relationships between different segments, which help to structure the text in a coherent manner. Centering Theory [Grosz et al. 1995] focuses on how discourse entities are managed across sentences, ensuring that the reader can follow the progression of ideas smoothly. These models have significantly influenced NLP research, particularly in the analysis and generation of coherent texts, offering insights into the mechanisms that make a text understandable and logically connected [Jurafsky and Martin 2024].

3. Related Work

As a central area of investigation in NLP, textual coherence, particularly in the context of local coherence, which focuses on the logical and sequential flow between adjacent sentences or paragraphs, has been extensively studied through models like the entity grid. Introduced by [Lapata and Barzilay 2005] and further developed by [Barzilay and Lapata 2008], the entity grid model abstracts a text into a grid that captures the distribution and transitions of discourse entities across sentences. By analyzing these patterns, the model can effectively infer the level of local coherence within a text. This approach has been widely adopted and has inspired numerous subsequent studies. For instance, [Elsner et al. 2007] enhanced coherence assessment by integrating the entity grid with a content model, while [Lin et al. 2011] refined the method by incorporating discourse relations, further advancing the field's understanding of how sentences connect and maintain coherence.

The shuffle test, introduced by [Barzilay and Lapata 2008], has become a standard method for evaluating local coherence models. This test involves comparing the coherence of a text in its original order versus a shuffled version, challenging models to recognize the coherent sequence. Studies like those by [Lin et al. 2011] and [Dias 2016] have used this test to validate the effectiveness of their models, highlighting its importance as a benchmark in coherence evaluation.

In contrast, global coherence, which concerns the overall unity and thematic consistency of a text, has received less attention but remains a key aspect for understanding how texts function as a whole. Early work by [Thompson 1986] emphasized the role of global coherence in enhancing readability and comprehension, arguing that a coherent text allows readers to follow the central theme or argument effortlessly. More recent contributions by [Sagi 2010] have explored the hierarchical structure of texts, demonstrating how well-organized discourse contributes to global coherence.

Recent advancements in NLP have introduced BERT and LLMs like GPT-3, which have significantly expanded the possibilities for coherence analysis. These models, trained on extensive datasets, exhibit a remarkable ability to capture both local and global coherence, leading to more refined and human-like assessments of textual structure. For example, [Braz Junior and Fileto 2021] applied BERT, specifically BERTimbau, in educational forums to measure coherence. By analyzing sentence embeddings, the model effectively assessed sentence order, accurately distinguishing between coherent and permuted texts, thus demonstrating its capability to capture nuanced textual relationships. Similarly, [Naismith et al. 2023] utilized GPT-4 for coherence assessment in educational contexts, where the model not only rated coherence but also provided explanatory rationales that closely aligned with human evaluations. This study demonstrated GPT-4's effectiveness in replicating human judgments and even surpassing traditional NLP metrics by offering rationale-supported evaluations, thereby highlighting its potential to enhance automated discourse coherence assessment and its applications in educational settings.

4. Methodology

The primary aim of this study is to analyze and compare the performance of various LLMs in evaluating textual coherence across different aspects using two distinct approaches: (i) through LLMs APIs and (ii) through LLMs chat interfaces To achieve this, we selected a diverse set of corpora for their relevance and variety in text types, which allows for a thorough assessment of the models' capabilities across different linguistic contexts. These datasets were preprocessed and, annotated as necessary to ensure consistency across the tasks.

One of the four corpora utilized in this study is the Corpus of Contemporary American English (COCA) [Davies 2008], which offers a balanced compilation of over one billion words across genres such as spoken language, fiction, academic texts, and web pages. For our analysis, we focused on the free portions of blog and academic sections of COCA, which together comprise a diverse range of coherence levels. The blog section, with 991 texts, is characterized by its informal and subjective nature, often exhibiting lower coherence, while the academic section, consisting of 256 texts, is known for its structured and precise language, typically demonstrating higher coherence. These sections were employed in both local and global coherence tasks, with specific subsets annotated for detailed global coherence analysis and incoherence identification.

The study also incorporates the CST News Corpus [Aleixo and Pardo 2008], which consists of 50 collections of Brazilian Portuguese news articles, each centered on a specific event or topic. Originally developed to support research on multi-document summarization, the corpus includes approximately 150 news articles and 300 human-generated summaries from various newspapers, such as Folha de São Paulo, Estadão, and O Globo. This diversity in sources makes the corpus particularly well-suited for coherence studies, as it allows for evaluating both local and global coherence in a multilingual context. The CST News Corpus was especially valuable for assessing model performance in Brazilian Portuguese, adding a multilingual dimension to our evaluations.

Another key corpus in this research is the Grammarly Corpus of Discourse Coherence (GCDC) [Lai and Tetreault 2018], which contains 4,800 texts from four real-world sources: Yahoo Answers, Clinton Emails, Enron Emails, and Yelp Reviews. Due to its context-dependent structure, the Yahoo Answers portion (1,200 texts) was excluded from our study. The Clinton Emails provide a mix of professional and personal correspondence, Enron Emails focus on formal business communication, and Yelp Reviews feature user-generated feedback on businesses. Each text is annotated for global coherence on a 3-point scale (low, medium, high), with 8,000 ratings from both expert and non-expert annotators via Amazon Mechanical Turk. These pre-existing annotations were used for comparing the models' performance against human judgments, enhancing our evaluation of global coherence tasks.

Lastly, the DDisCo corpus [Mikkelsen et al. 2022] was developed to fill a gap in resources for studying discourse coherence in Danish. It comprises 1,002 texts from two main sources: Reddit and Danish Wikipedia. The Reddit texts, totaling 501, consist of informal user-generated content, while the 501 Danish Wikipedia texts offer more formal, structured information. Each text is annotated for global coherence on a 3-point scale (low, medium, high) by linguistics experts. This corpus introduces linguistic diversity into our research, allowing us to evaluate model performance in another non-English context. It was particularly useful for assessing how well the models generalize across different languages and discourse structures.

4.1. Local Coherence Analysis

The local coherence analysis in this study employed the shuffle test, which evaluates text coherence by comparing the original order of sentences within each text to randomly shuffled version. This test was applied to texts from four corpora: COCA, CST News, GCDC, and DDisCo. A total of 2,318 texts were selected, comprising 991 blog texts and 256 academic texts from COCA, 251 news articles from CST News, 842 texts from GCDC, and 991 texts from DDisCo. Each text was segmented into sentences, and those containing fewer than four sentences were excluded as they would not allow for the 20 required permutations. The remaining texts were shuffled 20 times, generating 46,360 incoherent versions, resulting in a dataset of 48,678 texts for analysis.

The models' performance was evaluated using two distinct methods: (i) via LLMs APIs and (ii) through LLMs chat interfaces. In the API-based evaluation, texts were processed directly through automated API calls, streamlining the evaluation process. In contrast, the chat interface evaluation simulated real-world usage by submitting the texts through interactive prompts.

For both approaches, the models were provided with a standardized prompt for Local Coherence Analysis[1] to guide them in distinguishing between coherent and incoherent texts. Performance was measured using accuracy, precision, recall, and F1-score, comparing the models' classifications against the original text labels.

[1] https://github.com/bryankhelven/coherence-findings

4.2. Global Coherence Analysis

The global coherence analysis in this study aimed to evaluate the ability of various LLMs to assess the overall logical consistency and thematic organization of texts across a total of 2,142 texts. This analysis included 1,200 texts from the DDisCo Corpus and 842 texts from the GCDC Corpus, both of which already contained human annotations. Additionally, a new annotation phase was conducted for a subset of 100 texts from the COCA and CST News corpora, as these lacked pre-existing coherence labels. Three languages/linguistics experts evaluated this subset, which consisted of 10 academic texts and 60 blog texts from COCA, along with 30 news articles from CST News. For consistency, each text in this study was assigned a coherence score on a Likert scale ranging from low to high coherence (1 to 3), using the same scale previously adopted for assigning scores by the works of [Lai and Tetreault 2018] and [Mikkelsen et al. 2022]. This ensured that the evaluation of global coherence was standardized across the various corpora used in this analysis.

Following the annotation process, the study assessed the models' performance in global coherence tasks using two methods: (i) LLMs APIs for an automated process, and (ii) LLMs chat interfaces to simulate real-world, user-driven interactions. A total of 2,142 texts were used for this analysis, comprising the 100 manually annotated texts, 1,200 texts from the DDisCo corpus, and 842 texts from the GCDC corpus. Both methods utilized a standardized prompt for Global Coherence Analysis[2] to guide the models in assessing global coherence. The evaluation metrics were consistent with those used in the local coherence analysis, but in this case, the models' classifications were compared directly to the original human annotations (scores of 1, 2, or 3).

4.3. Incoherence Identification

The incoherence identification task evaluated the ability of various LLMs to detect segments within texts that disrupt logical flow. We used 130 texts for this task, including 100 texts previously annotated for global coherence (10 academic texts and 60 blog texts from COCA, 30 news articles from CST News) and an additional 30 texts from the GCDC corpus (10 each from Yelp, Clinton, and Enron). The same three annotators from the global coherence task identified incoherent segments, focusing on the categories of Incorrect Use of Logical Connectors, Unnecessary Repetition, Irrelevant Information, Contradictions, Sequence of Events, and Inconsistent Verb Tenses. Fleiss' Kappa, which scored 0.8326 and indicated excellent agreement, was chosen for its capacity to account for chance agreement among multiple annotators across various incoherence types. Each annotated segment was treated as a unit, ensuring robust reliability.

The annotators, familiar with each other, communicated freely to resolve difficulties, following a shared understanding of coherence from [Koch and Travaglia 2003]. The models' performance was evaluated using the same two methods as before – LLMs APIs and chat interfaces. However, in this task, each model was treated as an additional annotator. The agreement between model-generated annotations and human annotations was measured using Fleiss' Kappa to determine how closely the models aligned with human judgment. The prompt for incoherence identification[3] was also standardized and used across all models in this task.

[2] Available on GitHub (see first footnote).

[3] Ibid.

5. Results and Discussion

The results obtained during the execution of the analysis are summarized in Tables 1, 2, and 3, highlighting the performance of various models in both API and chat-based interactions.

Table 1. Performance Metrics for Local Coherence Classification

	API				Chat			
Model	**Acc**	**Pr**	**Re**	**F1**	**Acc**	**Pr**	**Re**	**F1**
Bard	0.756	0.755	0.740	0.748	0.739	0.742	0.739	0.740
Claude 3 Haiku	0.914	0.906	0.898	0.902	0.949	0.902	0.899	0.900
Claude 3 Opus	0.979	**0.991**	0.983	0.987	0.974	0.971	**0.973**	0.972
Claude 3.5 Sonnet	0.973	0.986	0.981	0.983	0.972	0.969	0.968	0.968
Gemini	0.978	0.989	0.980	0.985	0.971	0.971	0.970	0.970
GPT 3.5	0.918	0.908	0.901	0.905	0.962	0.905	0.902	0.903
GPT 4	0.970	0.982	0.980	0.981	0.969	0.966	0.965	0.965
GPT 4o	**0.982**	0.990	**0.988**	**0.989**	**0.977**	**0.975**	**0.973**	**0.974**
LLaMA 2 13b	0.831	0.825	0.816	0.820	0.888	0.821	0.818	0.819
LLaMA 2 7b	0.817	0.804	0.797	0.800	0.805	0.801	0.798	0.799

Table 2. Performance Metrics for Global Coherence Classification

	API				Chat			
Model	**Acc**	**Pr**	**Re**	**F1**	**Acc**	**Pr**	**Re**	**F1**
Claude 3 Haiku	0.959	0.918	0.921	0.920	0.911	0.871	0.875	0.875
Claude 3 Opus	**0.982**	**0.986**	**0.987**	**0.986**	**0.933**	**0.936**	**0.939**	**0.937**
Claude 3.5 Sonnet	0.980	0.984	0.982	0.983	0.930	0.934	0.931	0.932
Gemini	0.976	0.963	0.966	0.965	0.928	0.915	0.918	0.916
GPT 3.5	0.960	0.920	0.923	0.921	0.912	0.873	0.879	0.877
GPT 4	0.974	0.961	0.964	0.963	0.926	0.914	0.919	0.917
GPT 4o	0.978	0.965	0.968	0.967	0.930	0.918	0.920	0.919
LLaMA 2 13b	0.970	0.930	0.933	0.932	0.922	0.887	0.883	0.888
LLaMA 2 7b	0.968	0.928	0.931	0.930	0.920	0.881	0.884	0.883

Table 3. Fleiss' Kappa for Incoherence Identification

Model	**API**	**Chat**
Annotators only (baseline)	0.8326	0.8326
Claude 3 Haiku	0.7995	0.7653
Claude 3 Opus	0.8166	0.7987
Claude 3.5 Sonnet	0.8279	0.8082
Gemini	0.8119	0.7858
GPT 3.5	0.8038	0.7716
GPT 4	0.8152	0.8093
GPT 4o	**0.8316**	**0.8234**
LLaMA 2 13b	0.6787	0.6492
LLaMA 2 7b	0.5823	0.5418

Table 1 shows the performance metrics for Local Coherence Classification, with GPT 4o achieving the highest scores in both API and chat interactions. Claude 3 Opus and Claude 3.5 Sonnet also performed well, especially in the API interaction, which demonstrates their effectiveness in accurately identifying coherent texts. In contrast, LLaMA 2 13b and LLaMA 2 7b had similar lower performance on both scenarios, suggesting limitations in processing and classifying local coherence. Similarty, for Global Coherence Classification, GPT 4o and Claude 3 Opus stood out with the highest performance in both interaction modes, while Claude 3 Haiku had the lowest as shown in Table 2.

The results for the Incoherence Identification task are summarized in Table 3, where GPT 4o again demonstrated the highest agreement with human annotators, with a Fleiss' Kappa of 0.8316 in API interaction and 0.8234 in chat. Claude 3.5 Sonnet followed closely, with Kappa values of 0.8279 in API and 0.8082 in chat, while LLaMA models, particularly LLaMA 2 7b, showed significantly lower Kappa values, indicating that these models struggle more with identifying incoherent segments.

The difference in performance between API and chat interactions is notable, with all models generally performing better in the API-based tests across all scenarios. This may indicate that API interactions allow for more precise and structured processing, leading to higher accuracy and consistency.

6. Conclusions and Future Work

This study assessed the performance of LLMs in evaluating textual coherence at both local and global levels and identifying incoherences within various corpora. Models such as GPT 4o and Claude 3 consistently outperformed others, particularly in API-based evaluations, where they achieved high accuracy and reliability. In local coherence tasks, GPT 4o demonstrated an F1 score of 0.989 in API-based tests, while in global coherence tasks, Claude 3 Opus led with an F1 score of 0.986. However, chat-based interactions revealed a performance decline, with GPT 4o's F1 score dropping to 0.974 in local coherence and Claude 3 Opus to 0.937 in global coherence. This suggests that the mode of interaction impacts model effectiveness, with API-based methods being more stable.

Despite the strong performance of top models, the Incoherence Identification task proved challenging across the board. GPT 4o showed the highest agreement with human annotators (Fleiss' Kappa of 0.8316), but all models exhibited lower performance in chat-based settings. These findings underscore the need for improvement in this area, especially as lower-tier models like LLaMA 2 struggled significantly, with Fleiss' Kappa dropping as low as 0.5418 in chat-based evaluations.

These findings have practical implications for NLP as models like GPT 4o and Claude 3 can be integrated into proofreading tools, content generators, and educational software to improve textual coherence. Their ability to assess and enhance coherence benefits machine-generated content and helps users create cohesive texts. Recognizing the impact of interaction modes on performance guides developers in choosing effective deployment strategies, favoring API integrations for consistency and accuracy.

The study acknowledges threats to validity, particularly the risk that some of the evaluation corpora may have been part of the training data for the LLMs, potentially inflating performance. This overlap introduces biases that could compromise objectivity, as models may recall patterns from training instead of genuinely evaluating coherence. The assumption of coherence in original texts and the limited size and diversity of the annotated datasets also pose risks to the generalizability of the findings.

Future work should address these limitations by expanding the range of evaluated text types and incorporating larger, more diverse annotator groups, as well as utilizing new and manually collected corpus to ensure that the models have not had prior access to it. Additionally, exploring fine-tuning techniques and evaluating newer model architectures will be essential. The development of improved evaluation metrics and the exploration of cross-linguistic and multimodal coherence analysis are also recommended to enhance the robustness and applicability of LLMs in complex language tasks.

References

Aleixo, P. and Pardo, T. A. S. (2008). Cstnews: Um córpus de textos jornalísticos anotados segundo a teoria discursiva multidocumento cst (cross-document structure theory). Technical Report 326, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos-SP. 12p.

Barzilay, R. and Lapata, M. (2008). Modeling local coherence: An entity-based approach. In Knight, K., Ng, H. T., and Oflazer, K., editors, *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 141–148, Ann Arbor, Michigan. Association for Computational Linguistics.

Braz Junior, G. and Fileto, R. (2021). Investigating coherence in posts from a doubts forum in a virtual learning environment with bert. *Conference Paper*.

Charolles, M. (1978). *Introdução aos problemas da coerência dos textos: abordagem teórica e estudo das práticas pedagógicas*. Editora Pontes.

Davies, M. (2008). The corpus of contemporary american english (coca). Available online at https://www.english-corpora.org/coca/.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 4171–4186. Association for Computational Linguistics.

Dias, M. (2016). *Investigação de modelos de coerência local para sumários multidocumento*. PhD thesis, Universidade de São Paulo.

Elsner, M., Austerweil, J., and Charniak, E. (2007). A unified local and global model for discourse coherence. In Sidner, C., Schultz, T., Stone, M., and Zhai, C., editors, *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 436–443, Rochester, New York. Association for Computational Linguistics.

Freitas, A. R. P. (2013). *Análise automática de coerência usando o modelo grade de entidades para o português*. PhD thesis.

Grosz, B. J., Joshi, A. K., and Weinstein, S. (1995). Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.

Halliday, M. A. K. and Hasan, R. (1976). *Cohesion in English*. Longman.

Hoey, M. (2013). *Textual interaction: An introduction to written discourse analysis*. Routledge.

Jurafsky, D. and Martin, J. H. (2024). *Speech and Language Processing*, chapter 23. Draft, 3 edition. Accessed: 2024-02-29.

Koch, I. and Travaglia, L. (2003). *A coerência textual*. Editora Contexto.

Lai, A. and Tetreault, J. (2018). Discourse coherence in the wild: A dataset evaluation and methods. In *Proceedings of SIGdial*, pages 214–223.

Lapata, M. and Barzilay, R. (2005). Automatic evaluation of text coherence: models and representations. In *Proceedings of the 19th International Joint Conference on Artificial*

Intelligence, IJCAI'05, page 1085–1090, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Lin, Z., Ng, H. T., and Kan, M.-Y. (2011). Automatically evaluating text coherence using discourse relations. In Lin, D., Matsumoto, Y., and Mihalcea, R., editors, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 997–1006, Portland, Oregon, USA. Association for Computational Linguistics.

Mann, W. C. and Thompson, S. A. (1987). Rhetorical structure theory: Description and construction of text structures. In *Natural Language Generation*, pages 85–95. Springer Netherlands.

Mikkelsen, L. F., Kinch, O., Pedersen, A. J., and Lacroix, O. (2022). Ddisco: A discourse coherence dataset for danish. In *Proceedings of the 13th Language Resources and Evaluation Conference (LREC)*, pages 1234–1243.

Naismith, B., Mulcaire, P., and Burstein, J. (2023). Automated evaluation of written discourse coherence using gpt-4. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 394–403, Online. Association for Computational Linguistics.

Sagi, E. (2010). Discourse structure effects on the global coherence of texts.

Seno, E. R. M. and Rino, L. H. M. (2005). Co-referential chaining for coherent summaries through rhetorical and linguistic modeling. In *Proceedings of the Workshop on Crossing Barriers in Text Summarization Research/RANLP*, Borovets, Bulgaria. Núcleo Interinstitucional de Lingüística Computacional – NILC/USFCAR.

Thompson, I. (1986). Readability beyond the sentence: Global coherence and ease of comprehension. *Journal of Technical Writing and Communication*, 16(1):131–140.

Van Dijk, T. A. (1977). Text and context: Explorations in the semantics and pragmatics of discourse.