Automatic Measurement of Cohesion and Coherence in Agglutinative Languages: A Cohesion and Coherence Formula for Turkish

Muhittin Rahman Kalyoncu[1*], Batur Alp Akdoğan[2], Muhammet Memiş[3], Zehra Kamisli Ozturk[4]

[1] Hacettepe University, Graduate School of Educational Sciences, Ankara, Türkiye
[2] Eskişehir Technical University, Graduate School, Eskişehir, Türkiye
[3] Ondokuz Mayıs University, Faculty of Education, Samsun, Türkiye
[4] Eskişehir Technical University, Faculty of Engineering, Eskişehir, Türkiye
[*] Corresponding author's email: mrahmankalyoncu25@hacettepe.edu.tr / mrkalyoncu@sinop.edu.tr

DOI: https://doi.org/10.53482/2026_61_435

ABSTRACT

Due to the absence of an objective methodology and empirical research on measuring cohesion and coherence in Turkish, this study aims to establish a framework for the objective assessment of textual coherence and cohesion in this agglutinative language. To achieve this, the coherence, cohesion, and comprehensibility of 25 texts were first evaluated by 60 experts in language education. Next, structural text features, NLP-based similarity models, and large language models with potential to determine coherence and cohesion were identified and applied to the texts. The analyses highlighted the factors that can be employed in evaluating coherence and cohesion, and regression equations were constructed using these variables. Consequently, for the first time, objective and empirical methodologies have been developed to assess the coherence and cohesion of Turkish texts, while accounting for the structural characteristics of agglutinative languages.

Keywords: text, coherence, cohesion, agglutinative languages, Turkish

1 Introduction

Texts form the foundation of educational systems, academic scholarship, official practices, and everyday social life, and they are regarded as phenomena that people turn to and interact with across nearly all domains. The term *text* derives etymologically from the Latin *textus* (fabric), meaning "to weave" or "to knit". This association reflects the idea that, just as fabric is woven from threads to create a whole, a text emerges through the gradual and meaningful interconnection of its constituent elements. Accordingly, for a sequence of symbols or sounds, whether long or short, oral or written, to be recognized as a text, it must satisfy certain conditions, commonly referred to in the literature as the criteria of textuality.

In the first stage, the criteria of textuality, which are classified into text-centered and reader-centered categories, consist of seven elements: cohesion, coherence, intentionality, informativity, situationality, acceptability, and intertextuality (De Beaugrande and Dressler, 1981; Rahma et al., 2022). Among these, cohesion and coherence are considered text-centered criteria, whereas the others are reader-centered. Thus, for a discourse or a collection of symbols to be accepted as a text, it must first satisfy the criteria of cohesion and coherence, which constitute the integrity of a text (Aminovna, 2022; Diep and Le, 2024). Indeed, cohesion and coherence are the most frequently examined and emphasized textuality criteria compared to the others (McNamara et al., 2010; Jassim, 2023). Cohesion refers to the grammatical and lexical connections between sentences at the microstructural level, which emerge when the interpretation of certain elements in discourse depends on others (Halliday and Hasan, 1976; Dascalu et al., 2015; Kouti, 2022). Coherence, by contrast, represents the relationships and patterns that should exist among textual elements in terms of meaning and logical integrity (Davoodi and Kosseim, 2017; Karatay, 2010; van Dijk, 1981). Accordingly, cohesion concerns the interpretation of the connections among words, components, and conveyed ideas in a sociocultural context (McNamara et al., 2002), whereas coherence reflects the psychological processes underlying the mental representations that individuals actively construct as they attempt to understand a text (Graesser et al., 2007; Sanders and Maat, 2006). In this sense, cohesion can be described as surface-level linguistic relations, that is, the micro level, while coherence pertains to the semantic integrity of a text, namely the macro level (De Haas et al., 2000). Despite the complex and debated relationship between them (Thompson, 1994), cohesion and coherence can be regarded as inseparable and mutually reinforcing (Bui et al., 2021; Lismay, 2020).

The levels of cohesion and coherence in texts directly affect their quality and, consequently, shape the ways in which readers interact with them. Research in this area has shown that the degree of cohesion and coherence influences text comprehension (Gasparintou and Grigoriadou, 2013; Graesser et al., 2003; Hall et al., 2014; Ozuru et al., 2009; Reed and Kershaw-Herrera, 2016), text production (Gómez González, 2013), the strategies used in text comprehension (Follmer and Sperling, 2018), writing quality (Collins, 1998), the formation of mental representations (O'Reilly and McNamara, 2007), as well as the persuasiveness of texts, readers' attitudes toward them, and the retention of information (Kaakinen et al., 2011). Moreover, examining the degree of cohesion makes it possible to distinguish between texts written in a native language and those written in a foreign language (Azadnia et al., 2016; Crossley and McNamara, 2009). To manage these wide-ranging effects of cohesion and coherence, it is essential to measure them first.

In the international literature, various methodologies have been developed to measure cohesion, and several applications for its automatic assessment have been produced. The earliest approaches relied on vector-based similarity models, most notably LSA (Foltz et al., 1998). Over time, methodological diversity increased as structural features of texts – such as connectives, synonyms, and pronouns – were

incorporated into vector-based similarity models alongside natural language processing (NLP) techniques such as TF-IDF, LDA, and BERT. For instance, Crossley et al. (2016) developed the first version of the TAACO model using variables such as TTR, synonyms, connectives, and pronouns. They later extended it to TAACO 2.0 by integrating NLP models such as LSA, LDA, and word2vec (Crossley et al., 2019). Similarly, Dascalu et al. (2015) created cohesion measurement tools by combining structural text features with NLP models. More recently, research on the automatic assessment of textual cohesion and coherence (He et al., 2024; Ueda, 2021; Zhao et al., 2023) has advanced through the application of increasingly diverse NLP methods. Nevertheless, because each language has unique characteristics, these methodologies remain specific to the languages for which they were originally designed.

When it comes to Turkish, no empirical study has examined the statistical factors influencing the cohesion of texts, and no objective methodology has been proposed for its automatic measurement. Although many studies in Turkish literature address cohesion and coherence, they generally rely on rubric-based scoring keys or checklist-type scales, and in some cases, texts are evaluated intuitively.

2 Current study

Given the reasons outlined above, it is essential to measure the cohesion and coherence of texts, which serve as key evaluation criteria in academic, daily, and social contexts, in an objective and practical way for Turkish and to develop a methodology for this purpose. In this study, a set of texts with varying levels of cohesion and coherence was first presented to experts in language education, who provided objective evaluations of their cohesion and coherence. Following this stage, structural features with the potential to affect cohesion and coherence, together with NLP-based similarity models widely used in the international literature, were examined and applied to each text in the sample.

Following the described procedures, the structural features and NLP models associated with the cohesion and coherence levels of the texts were identified through correlation analysis, and regression analyses were conducted using these variables. In this way, a methodology for measuring cohesion and coherence in Turkish was developed. In addition, the study examined the performance of rapidly evolving artificial intelligence tools and large language models in determining the coherence and cohesion of Turkish texts.

The findings obtained in this process are expected not only to contribute to the development of a methodology for Turkish but also to support studies in other agglutinative languages. This is because the structural characteristics of agglutinative languages set them apart from other language types and require specific preprocessing steps in analyses. In this context, the methodology developed for Turkish as an agglutinative language is anticipated to offer significant advantages for research conducted in other agglutinative languages.

Accordingly, the present study aimed to develop an objective and practical methodology for Turkish that accounts for the structural characteristics of agglutinative languages and to examine the functionality of existing resources in determining the levels of coherence and cohesion in texts. In line with this aim, the following research questions were addressed:

1. Which structural text features influence the cohesion and coherence of texts?

    1.1. According to the participants in the study group, what are the coherence and cohesion values of the texts in the sample?

    1.2. What structural text features may be related to the coherence and cohesion of texts?

    1.3. What is the status of the structural text features that may affect the cohesion and coherence of texts for the texts included in the sample?

    1.4. What is the relationship between structural text features and the coherence/cohesion of texts?

2. Can NLP-based text similarity models be used to measure the coherence and cohesion of texts?

    2.1. Which NLP-based text similarity models can be used to measure text coherence and cohesion?

    2.2. What are the values obtained from NLP-based text similarity models for the texts in the sample?

    2.3. What is the relationship between the values obtained from NLP-based text similarity models and the coherence and cohesion of the texts in the sample?

3. What are the forms of the regression equations to be constructed with the variables related to text coherence/cohesion?

4. Do artificial intelligence models provide valid results in determining text coherence/cohesion?

3 Methodology

3.1 Research design

This study employed a mixed-methods approach, combining both qualitative and quantitative research methods, and utilized the convergent design, which is one type of this approach. The purpose of the convergent design is to integrate the results obtained from the analysis of qualitative and quantitative data. Such integration provides different perspectives on both types of data, enabling the problem to be defined from both quantitative and qualitative dimensions and allowing it to be examined from multiple viewpoints through the merging of these data (Creswell, 2017).

The qualitative phase of this research involved examining the texts in the sample in terms of their structural features and evaluating them through various software tools, while the quantitative phase consisted of collecting the opinions of a study group of 60 participants regarding the texts in the sample. The integration phase entailed combining the qualitative evaluations and quantitative data on the texts and

analyzing them through correlation and regression analyses. These processes are summarized in Figure 1 below.

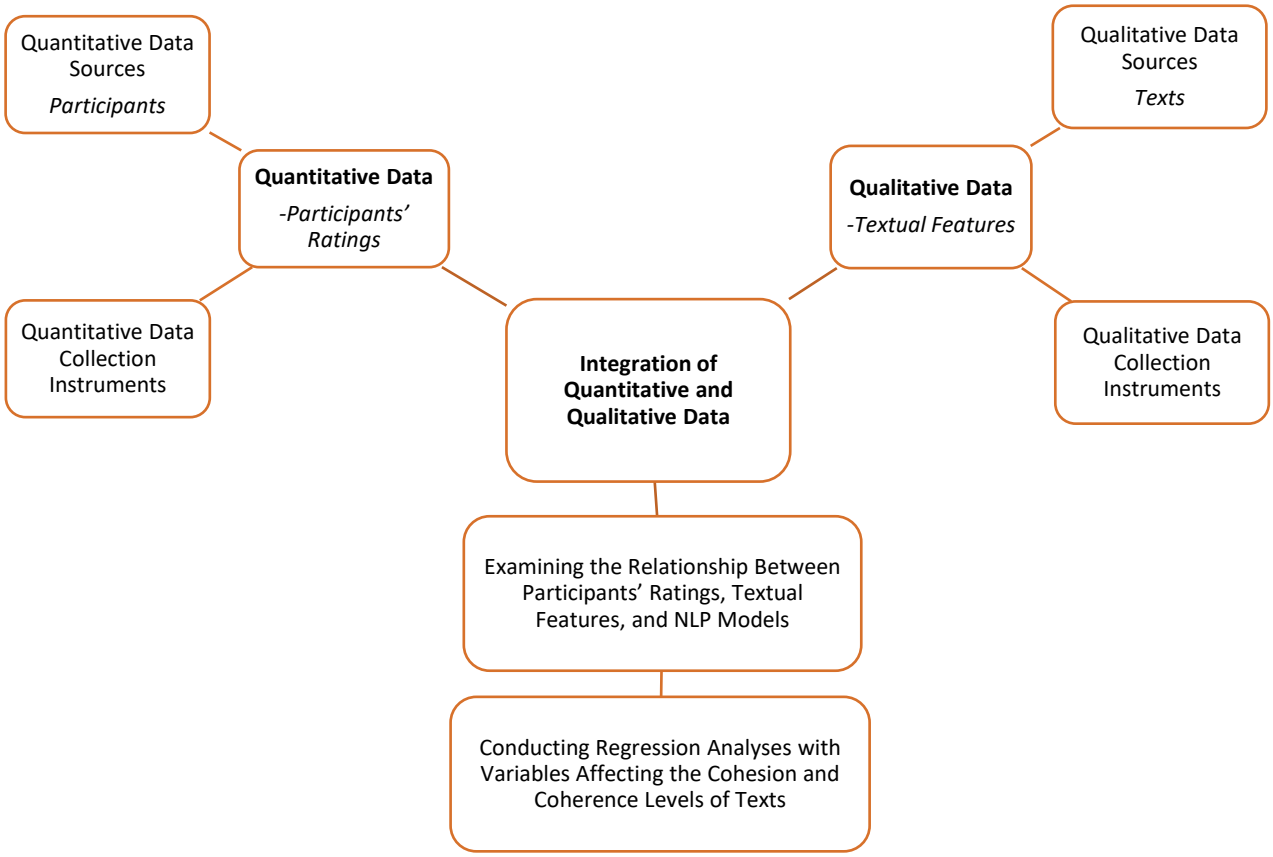

Figure 1: Research design

In addition to the activities summarized in the diagram above, the study also aimed to evaluate the effectiveness of artificial intelligence tools in detecting the cohesion and coherence of texts. For this purpose, the texts were also presented to AI tools, and the correlation between the values obtained and the assessments of the study group was examined. These procedures constituted the second phase of the research and reflected the characteristics of a correlational study.

3.2 Sample/study group

This research utilized two different data sources: texts and participants. The sample of the study consisted of 25 texts selected or created through the extreme case sampling method. The reason for choosing extreme case sampling in the selection/creation process was the assumption that phenomena with opposing characteristics would provide a clearer opportunity to observe variability (Büyüköztürk et al., 2023). In this context, 25 texts with differing levels of cohesion and coherence and distinct structural features were selected or constructed by the researchers.

In designing the sample, three approaches were employed: directly using existing texts, disrupting the coherence and cohesion of existing texts, and producing incoherent, non-cohesive texts. Accordingly, 10 of the texts in the sample were directly taken from various sources, 9 were taken from different sources and then had their coherence and cohesion deliberately disrupted, and 6 were constructed to be inherently lacking in coherence and cohesion. In the processes of disrupting textual coherence and cohesion and generating incoherent, non-cohesive texts, the researchers received support from the O3 model of the ChatGPT AI tool. First, the AI tool was instructed to disrupt the cohesion of the original texts, and appropriate levels of cohesion and coherence were achieved by revising the outputs by the researchers. No control mechanism was implemented in this process; the texts were structured based on the researchers' intuition and experience.

After the texts were created, a study group was formed to objectively determine their levels of coherence and cohesion. For this process, purposive sampling was employed to recruit participants with specific expertise in Turkish language education who could provide the most accurate evaluations of the texts' coherence and cohesion. Based on this requirement, a study group of 60 volunteers was assembled, representing different stakeholders in Turkish language education. The group included 2 academics working in the field, 2 Turkish language teachers, 10 graduate students specializing in Turkish language education, and 46 undergraduate students enrolled in a Turkish language teaching program.

3.3 Data collection instruments

Different data collection instruments were employed in this study. First, a form was developed by the researchers to gather the opinions of the participants in the study group regarding the texts in the sample. This form required participants to make judgments about the cohesion and coherence of the texts. Specifically, each participant was asked to answer four questions for each text. These questions involved the categorical assessment of text coherence, as well as the numerical evaluation of coherence, comprehensibility, and cohesion on a scale from 1 to 10. The questions are presented below:

"1 – Categorical assessment of the text's coherence

A – Completely Coherent B – Coherent C – Incoherent D – Completely Incoherent

2 – Numerical evaluation of the text's coherence on a scale from 1 to 10

(1 = Completely Incoherent, 10 = Completely Coherent)

3 – Numerical evaluation of the text's comprehensibility on a scale from 1 to 10

(1 = Completely Incomprehensible, 10 = Completely Comprehensible)

4 – Numerical evaluation of the text's cohesion on a scale from 1 to 10

(1 = Completely Non-cohesive, 10 = Completely Cohesive)"

The above questions were repeated for each text in the form, and accordingly, participants answered them 25 times. Following the development of the form to be administered to the study group, the "Coherence and Cohesion Elements Form" was designed by the researchers to identify the structural features of the texts in the sample. In the process of creating this form, the potential factors affecting text coherence and cohesion were first determined by the researchers, with reference to the literature, as TTR, reference ratio, proportion of cohesive devices, proportion of conjunctions, pronoun ratio, demonstrative adjective ratio, and preposition ratio. A form was then created to detect these structural features in the texts.

After identifying the structural features that may influence the cohesion and coherence of texts, NLP-based text similarity models that could be used to determine cohesion and coherence were considered. In this context, it was considered that the prominent vector-based similarity approaches in the field, TF-IDF, MiniLM, BERTurk, and FastText, were functional in assessing text cohesion and coherence. The characteristics of these NLP-based text similarity models are explained below:

TF-IDF

TF-IDF is a statistical measure that quantifies the relative importance of a word within a document. It is based on the product of the term frequency (TF) of a word in the document and the inverse document frequency (IDF), which indicates how distinctive the word is across the entire collection. In this study, each sentence was treated as an independent "document," TF-IDF vectors of the sentences were extracted, and the average cosine similarity between consecutive sentences was calculated to obtain the coherence score.

Multilingual BERT

Multilingual BERT is a context-sensitive language model pre-trained on more than one hundred languages. It evaluates words not only in isolation but also in relation to their interactions with other words in the sentence. In practice, a semantic vector (embedding) was generated for each sentence, and the average cosine similarity between adjacent sentence vectors was used as the coherence score of the text.

BERTurk

BERTurk is a BERT model specifically trained on a large and diverse Turkish corpus. Due to its sensitivity to the morphological characteristics of Turkish, it is expected to perform effectively in Turkish

text-related tasks. In this study, sentences were converted into vectors using BERTurk, and the average cosine similarity between consecutive sentence pairs was calculated to obtain the coherence value.

FastText

FastText is a library designed for learning word representations, and it vectorizes words based on character n-grams. For sentence representation, a single sentence vector is obtained by averaging the word vectors that constitute the sentence. Using these vectors, the average cosine similarity between adjacent sentences was calculated, and the coherence of the texts was measured in this way.

After identifying the software used for text similarity detection that could be functional in measuring text coherence and cohesion, artificial intelligence tools that could be included in the study were considered in order to evaluate their performance in this task. In this context, globally prominent AI tools such as ChatGPT, DeepSeek, Grok, Gemini, and Perplexity were employed. During the research process, the O3 model was used within ChatGPT, the 2.0 Flash model within Gemini, the Grok 3 model within Grok, the V3 model within DeepSeek, and the Claude 3 Sonnet model within Perplexity. All evaluations were conducted via publicly accessible web interfaces rather than API access; therefore, no control over model configuration or sampling hyperparameters was possible, and all outputs were generated using the default settings implemented by each platform.

3.4 Data collection and analysis

In the data collection process, the texts in the sample were first presented to the participants in the study group through the form developed by the researchers. The data were collected in 2025 using Google Forms. After gathering the participants' evaluations, document analysis was conducted to determine the structural features of the texts, and the demonstrative adjective, pronoun, conjunction, preposition, reference, cohesive device, and TTR values of the texts in the sample were identified. In addition to these values, two new variables were created during the process: "ratio of demonstrative adjectives to pronouns" and "ratio of conjunctions to pronouns". For the calculation of these variables, the number of the relevant items was divided by the total number of words. However, in determining TTR ratios for agglutinative languages, lemmatization was applied prior to the calculation of the respective values. Moreover, since the texts in the sample were generally similar in length and none exceeded 500 words, MATTR (Covington and McFall, 2010), a more recent approach that compensates for the disadvantages of TTR, was not employed. In future studies involving longer texts, the use of MATTR would be more appropriate.

After obtaining the study group's evaluations of the texts and identifying their structural features, NLP-based text similarity models were employed, and the similarity values of the texts were determined using the Python programming language. In the analysis process, the extent to which a text progressed coherently at the sentence level was evaluated through a numerical approach. The texts were first segmented into sentences, a vector representation was generated for each sentence, and the similarities

between consecutive sentence pairs were calculated. The average of these similarities was then interpreted as the coherence score of the text. Explanations of this process were presented in paragraphs consistent with the flow of the code.

Sentence segmentation and preprocessing

Sentence boundaries were determined using Zemberek's *TurkishSentenceExtractor* component, which is sensitive to the rules of Turkish. Access to Zemberek from Python was provided through JPype1. In the preprocessing stage, four versions of each text were prepared:

1. In the raw version, no intervention was made to the text, and the models' responses to noisy data were observed.

2. In the stemmed version, stemming was applied with Zemberek's *TurkishMorphology* component, lemmatization was applied, and morphological diversity was reduced.

3. In the cleaned raw version, elements such as punctuation and numbers were removed with simple regular expressions, and frequently occurring functional words were eliminated using NLTK's Turkish stopword list.

4. The cleaned stemmed version combined both the stemming and cleaning steps.

These variants allow for a comparison of the sensitivity of the following representation methods to morphology and textual noise.

Vectorization with TF-IDF

As a statistical baseline, TF-IDF was implemented using scikit-learn's *TfidfVectorizer* class. Tokenization was carried out at the word level, and the Turkish stopword list was activated through the *stop_words* parameter when necessary. Each sentence was treated as an independent document, term weights were calculated, and sentence vectors were obtained. The method was applied to each of the preprocessing versions 1–4.

BERTurk

The Turkish-specific BERTurk model (*dbmdz/bert-base-turkish-cased*) was implemented using the *transformers* and *torch* frameworks. For input preparation, the tokenizer was automatically loaded with *AutoTokenizer.from_pretrained(...)*, and WordPiece rules were applied. Sentence representations were obtained using the output vector corresponding to the [CLS] token; in this process, no additional mean pooling step was included. Since the [CLS] vector is expected to capture the holistic meaning of the sentence, this representation was directly used as the embedding.

Multilingual MiniLM

For multilingual scenarios, a MiniLM-based model (e.g., *paraphrase-multilingual-MiniLM-L12-v2*) within the *sentence-transformers* framework was employed. In this setup, the tokenizer was automatically loaded by Hugging Face's *AutoTokenizer*, taking into account the subword rules used by the underlying model. Since the framework manages tokenization and vector generation end-to-end, sentence embeddings were obtained directly. As no additional pooling configuration was specified, the sentence vectors returned by the model were used as is. The lightweight architecture and multilingual pretraining of MiniLM provide fast and reliable representations across different language inputs.

FastText

In the FastText approach, words are vectorized based on character n-gram components. This feature enables the generation of consistent vectors even for out-of-vocabulary words. Tokenization was performed using simple whitespace-based word splitting, with light regex cleaning applied when necessary. Sentence representations were created by calculating the arithmetic mean of the word vectors within each sentence. The same procedure was repeated for all four preprocessing versions: in the raw versions, natural noise was preserved, while in the cleaned and/or stemmed versions, noise and morphological variation were reduced, resulting in more balanced averages.

Similarity measurement and scoring

In this study, coherence calculation was based on measuring the semantic similarity between adjacent sentences. Previous English-focused studies have reported that text coherence can be assessed using cosine similarity between consecutive sentence vectors. This approach was adapted to Turkish in part of the study, where sentence representations were generated using pre-trained language models and methods such as TF-IDF.

In this study, cosine similarity was used as the numerical measure of text coherence and cohesion. Cosine similarity considers the orientation of vectors and is independent of their magnitude. Its value ranges between −1 and 1; however, for representations without negative components, such as TF-IDF, it is practically limited to the range of 0 to 1. For two sentence vectors A and B, it is calculated as cos(θ) $\cos\theta = \frac{A \,.\, B}{\|A\| \, \|B\|}$. Similarity calculations were performed using this formula. For each text, the computation was carried out on consecutive sentence pairs (v_i , v_{i+1}); cosine values were obtained for all pairs, and their arithmetic mean was reported as the coherence score of the text. These scores take values between 0 and 1.

Summary of library roles

In the process, Zemberek and JPype1 were used for sentence segmentation and stemming; NLTK for providing the Turkish stopword list; scikit-learn for TF-IDF vectorization and cosine similarity calculation; *transformers* and *torch* for running the BERTurk model and its tokenizer (*AutoTokenizer*); *sentence-transformers* for generating MiniLM-based sentence embeddings; FastText for providing word

vectors and applying the sentence-level averaging approach; and NumPy and Pandas for data handling and numerical operations. This structure enabled the seamless integration of steps within the code flow and ensured the reliable generation of coherence scores.

After obtaining the similarity values of the texts, their cohesion and coherence were also evaluated using artificial intelligence tools. In this process, all tools were provided with a fixed instruction, after which they were asked to evaluate the texts presented to them according to this instruction. Indeed, on April 29, 2025, the content of the command given to all AI tools was identical to the questions included in the form administered to the participants of the study group.

After completing the processes described above, the analysis phase was initiated. In this phase, the mean values of the numerical evaluations made by the study group on the texts were first calculated to determine the levels of cohesion and coherence. Subsequently, correlation analysis was employed to examine the relationships between text coherence levels and the other variables under consideration. This procedure was carried out separately for each data group: first, the relationship between the structural features of the texts and their coherence; then, the correlation between NLP-based text similarity models and coherence; and finally, the correlation between the assessments of AI tools and those of the study group.

After the correlation analyses were completed, variables closely associated with text coherence and cohesion were identified, and multiple linear regression analyses were conducted with the appropriate ones. With the completion of the regression analyses, the data analysis process was concluded. SPSS 25 was used for all normality, correlation, and multiple linear regression analyses. Validity and reliability studies related to this process are presented in the following section.

3.5 Validity and reliability

Within the scope of validity and reliability studies conducted during the research process, normality analyses were first performed. Examination of the results revealed that the comprehensibility evaluation scores of the texts provided by ChatGPT and the pronoun ratios in the texts exceeded the ± 2 threshold. On the other hand, all remaining data fell within the ± 2 range and exhibited a normal distribution (George and Mallery, 2010). To ensure the integrity of the research process, data that did not display a normal distribution were excluded from the study, and analyses were continued with the data that did.

After calculating the normality values for the correlation analyses, the correlations among the variables to be included together in the regression analysis were examined. As is well known, another requirement for a valid and reliable regression analysis is that there should not be high correlations among the independent variables (Can, 2020). In this context, when correlations at the level of 0.80–0.90 are observed between variables, multicollinearity can be said to exist, and such variables should not be included in the analysis together (Büyüköztürk, 2005; Can, 2020). Accordingly, the correlations among the variables to be jointly considered in the regression analyses are presented in Table 1 below:

Table 1: Correlations among variables to be used in regression analyses.

	V	Mean	Std. Dev.	1	2
1	TF-IDF[4]	0.07	0.05	1	
2	BERTurk[3]	0.83	0.10	–0.616**	1
3	Conjunction and Pronoun Ratio	0.06	0.03	0.456*	–0.498*

*N = 25; **p < 0.01; *p < 0.05*

An examination of Table 1 shows that none of the variables included in the multiple regression analyses had correlations higher than 0.8. In this regard, it can be stated that the simultaneous use of these variables would not cause any multicollinearity problems.

In multiple regression analyses, while it is necessary that the variables included in the analysis do not exhibit high correlations, it is also important that certain values obtained as a result of the analysis fall within acceptable ranges. Information regarding these values is presented in Table 2 below:

Table 2: Analyses related to the reliability of the regression analysis.

Regressions	Cooks		Std. Res.		Durbin	V	Collinearity Statistics	
	Min	Max	Min	Max			Tolerance	VIF
1	0.00	0.151	–1.613	1.894	2.292	TF-IDF[4]	0.591	1.693
						BERT[3]	0.561	1.782
						Conj. and Pron. R.	0.716	1.396
2	0.00	0.307	–1.478	1.579	1.479	TF-IDF	0.620	1.612
						BERT	0.620	1.612

An examination of Table 2 shows that the results of the analyses conducted in the study are presented. First, regarding the Variation Inflation Factor (VIF), it was found that all variables had values below 2. Since a VIF value lower than 5 indicates the absence of a multicollinearity problem in regression analysis (Craney and Surles, 2002; Güçlü, 2020; Ringle et al., 2015), this criterion was satisfied. Similarly, the "Tolerance" value should be greater than 0.2 (Gujarati, 2004), and all tolerance values in the Table 2 were within the acceptable range. In addition, the Durbin-Watson coefficient, which tests for autocorrelation (Durbin and Watson, 1971), takes values between 0 and 4, with values close to 2 being desirable (Ersöz and Ersöz, 2019). Furthermore, the "Cook's" statistic in the Table 2 relates to outliers, and the fact that none of the values exceeded 1 indicates that there was no outlier problem (Cook and Weisberg, 1982). In this regard, it can be stated that the analyses carried out did not reveal any issues related to multicollinearity or outliers.

In addition to the analyses conducted in the study, the reliability of the data collected from the participants was also examined. In this context, an intraclass correlation analysis was performed to determine the degree of consistency and reliability of the ratings assigned by the expert participants for the

dimensions of coherence, cohesion, and comprehensibility of the texts. Detailed results of this analysis are presented in Table 3 below:

Table 3: Consistency of the expert group's ratings.

V	N	Model	Type	Measure	ICC	95% CI	Interpretation
Coherence	25	Two-Way Random	Absolute Agreement	Average Measures	0.991	[0.98, 0.99]	High Reliability
Cohesion	25	Two-Way Random	Absolute Agreement	Average Measures	0.978	[0.96, 0.98]	High Reliability
Comprehensibility	25	Two-Way Random	Absolute Agreement	Average Measures	0.985	[0.97, 0.99]	High Reliability
Total	75	Two-Way Random	Absolute Agreement	Average Measures	0.985	[0.98, 0.99]	High Reliability

Examination of Table 3 indicates a high level of consistency among the experts' ratings across all evaluated dimensions. In this regard, it can be concluded that the ratings provided by the participants comprising the expert group in the study are reliable.

3.6 Ethics committee approval

The data collection procedure defined within the scope of this research and the data collected were confirmed to present no ethical issues and were approved by the Ethics Committee of Social and Human Sciences at Sinop University on May 9, 2025, with decision number 209-310. Participation in the study was entirely voluntary, no personal data were collected from the participants, informed consent was obtained, participants were informed that they could withdraw from the study at any time without any consequences, and all ethical principles were adhered to throughout the research process.

4 Results[1]

In the research process, the cohesion, coherence, and comprehensibility values of the texts in the sample were first determined based on the evaluations of the study group, and subsequent analyses were carried out regarding these assessments. The findings obtained from this analysis are presented in Table 4 below:

Table 4: Relationships among the coherence, cohesion, and comprehensibility of the texts in the sample.

	V	M	Std. Dev.	1	2
1	Coherence	6.09	2.54	1	
2	Comprehensibility	6.51	2.26	0.975**	1
3	Cohesion	6.56	1.99	0.950**	0.976**

*N = 25; **p < 0.01*

[1]Analyses related to research questions 1.1, 1.3, and 2.2 are available at the link provided in the *Data Availability Statement* section.

An examination of Table 4 indicates that the expert group's evaluations of the texts' coherence, cohesion, and comprehensibility were highly consistent and similar. In this regard, it can be stated that these elements almost point to the same phenomena. Accordingly, in the subsequent stages of the research process, the concepts of coherence and cohesion were considered together in the evaluations.

After determining the levels of cohesion and coherence, the texts in the sample were analyzed in terms of their structural features to address the question, "What is the relationship between structural text features and the coherence/cohesion of texts?" In this process, the relationship between the structural features of the texts and the evaluations of the study group participants was examined, and the analysis was carried out. The relevant findings are presented in Table 5 below:

Table 5: Relationship between the structural features of texts and their coherence.

	V	Mean	Std. Dev.	1	2	3	4	5	6	7	8
1	Coherence	6.09	2.54	1							
2	Conjunctive Cohesion	0.01	0.01	0.480*	1						
3	Conjunction Ratio	0.05	0.03	0.621**	0.693**	1					
4	Reference Ratio	0.01	0.02	0.679**	0.533**	0.379	1				
5	Preposition Ratio	0.01	0.01	0.124	–0.028	0.005	–0.039	1			
6	Conj. and Pron. Ratio	0.06	0.03	0.694**	0.519**	0.908**	0.480*	–0.043	1		
7	TTR	0.45	0.13	–0.537**	–0.416*	–0.426*	–0.384	0.112	–0.366	1	
8	Dem. Adj. R.	0.01	0.01	0.664**	0.680**	0.549**	0.893**	0.131	0.585**	–0.398*	1
9	Dem. Adj. and Pron. R.	0.02	0.02	0.646**	0.273	0.373	0.743**	0.137	0.646**	–0.245	0.779**

*N = 25; **p < 0.01; *p < 0.05*

An examination of Table 5 shows that, with the exception of the preposition ratio, all of the structural features considered in this study had a significant relationship with the coherence of the texts. The strength of these relationships was at a moderate level.

After determining the relationship between the structural features of the texts and their coherence/cohesion, the texts in the sample were analyzed using NLP-based text similarity models to address the question, "To what extent are NLP-based text similarity models related to the coherence and cohesion of texts?" The findings of this analysis are presented in Table 6 below.

An examination of Table 6 shows that the methodologies employed in applying the models influenced the relationship between the models and text coherence/cohesion. Specifically, for the MiniLM model, the highest correlation was obtained in the analysis conducted with the raw text (r = 0.733); for the BERTurk model, the highest correlation was achieved in the analysis where the words in the text were stemmed (r = 0.722); and for the TF-IDF model, the highest correlation was observed when the words were stemmed and punctuation marks were removed (r = 0.764). It is also noteworthy that all three of these models exhibited strong correlations with text coherence/cohesion. On the other hand, analyses

conducted with the FastText model did not yield correlations as high as those obtained from the other models.

After determining the relationship between text similarity software, structural features, and the coherence/cohesion of texts, the question “What are the forms of the regression equations constructed with

Table 6: Relationship between NLP-based text similarity models and text coherence.

	V	M.	Std. D.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	Coherence	6.10	2.54	1															
2	MiniLM[1]	0.39	0.15	0.733**	1														
3	MiniLM[2]	0.42	0.15	0.726**	0.994**	1													
4	MiniLM[3]	0.40	0.15	0.647**	0.976**	0.971**	1												
5	MiniLM[4]	0.43	0.14	0.655**	0.973**	0.976**	0.993**	1											
6	BERTurk[1]	0.81	0.12	-0.633**	-0.571**	-0.595**	-0.536**	-0.564**	1										
7	BERTurk[2]	0.71	0.12	-0.424*	-0.779**	-0.801**	-0.803**	-0.814**	0.527**	1									
8	BERTurk[3]	0.83	0.11	-0.722**	-0.723**	-0.728**	-0.664**	-0.664**	0.771**	0.634**	1								
9	BERTurk[4]	0.70	0.12	-0.613**	-0.601**	-0.615**	-0.520**	-0.521**	0.448*	0.553**	0.658**	1							
10	TF-IDF[1]	0.03	0.03	0.693**	0.698**	0.669**	0.659**	0.633**	-0.259	-0.411*	-0.505*	-0.422*	1						
11	TF-IDF[2]	0.03	0.03	0.678**	0.692**	0.662**	0.658**	0.627**	-0.250	-0.404*	-0.492*	-0.403*	0.997**	1					
12	TF-IDF[3]	0.07	0.05	0.761**	0.775**	0.766**	0.751**	0.729**	-0.463*	-0.551**	-0.613**	-0.517**	0.899**	0.897**	1				
13	TF-IDF[4]	0.07	0.05	0.764**	0.779**	0.770**	0.755**	0.733**	-0.465*	-0.553**	-0.616**	-0.518**	0.900**	0.897**	0.999**	1			
14	FastTexT[1]	0.59	0.09	0.323	0.691**	0.656**	0.678**	0.645**	-0.121	-0.614**	-0.324	-0.316	0.614**	0.618**	0.612**	0.612**	1		
15	FastTexT[2]	0.62	0.09	0.310	0.708**	0.678**	0.707**	0.676**	-0.151	-0.660**	-0.340	-0.323	0.639**	0.643**	0.643**	0.641**	0.988**	1	
16	FastTexT[3]	0.67	0.07	0.499*	0.786**	0.758**	0.781**	0.749**	-0.270	-0.651**	-0.445*	-0.486*	0.635**	0.637**	0.709**	0.707**	0.892**	0.895**	1
17	FastTexT[4]	0.61	0.09	0.460*	0.753**	0.721**	0.749**	0.716**	-0.243	-0.649**	-0.435*	-0.442*	0.683**	0.684**	0.729**	0.727**	0.931**	0.937**	0.972**

*N = 25; **p < 0.01 *p <0.05*

[1] *= Analysis conducted with the raw text*

[2] *= Analysis conducted with the punctuation-cleaned raw text*

[3] *= Analysis conducted with the stemmed text*

[4] *= Analysis conducted with the stemmed and punctuation-cleaned text*

the variables related to text coherence/cohesion?" was addressed, and analyses were carried out. In this process, potential combinations were first considered, and the most meaningful and functional model was sought by taking into account the correlations among variables, their ease of calculation, and their joint variance explanation rates. Accordingly, two different combinations were identified, and analyses were conducted on this basis. The first analysis performed in this context is presented in Table 7 below:

Table 7: First regression analysis.

V	*b*	*SE*	95% CI		*t*	*p*
			LL	*UL*		
Constant	8.489	3.175	1.887	15.091	2.674	0.014
TF-IDF[4]	21.124	6.586	7.429	34.820	3.208	0.004
BERTurk[3]	–6.654	3.275	–13.465	0.157	–2.032	0.055
Conj. and Pronoun Ratio	23.801	8.045	7.070	40.533	2.958	0.008
$R^2 = 0.78$	$F_{(3-21)} = 24.560$	$p < 0.001$	N = 25			

Table 7 presents the first regression analysis conducted in the research process. In this analysis, cosine similarity results calculated using the TF-IDF4 and BERTurk3 models, which did not show high correlations with each other, were employed, and the ratios of conjunctions and pronouns were also included as variables. Examination of the results shows that the model explained 78% of the variance in text coherence/cohesion. In addition, the model was found to be significant as a result of the regression analysis ($F[3 – 21] = 24.560$, $p < 0.01$). Accordingly, the regression equation that can be established for text coherence/cohesion is as follows:

$$\text{Text Coherence/Cohesion} = (8.489) + (21.124[X_1]) + (–6.654[X_2]) + (23.801[X_3])$$

X_1= TF-IDF[4] X_2= BERTurk[3] X_3= Conj. and Pronoun Ratio

After establishing the first regression equation, a second analysis was conducted in which only vector-based similarity models were used as variables, with the aim of enabling the direct assessment of text cohesion/coherence without the need for any prior structural examination of the text. The details of this analysis are presented below:

Table 8: Second regression analysis

V	*b*	*SE*	95% CI		*T*	*p*
			LL	*UL*		
Constant	12.318	3.371	5.326	19.309	3.654	0.001
TF-IDF[4]	25.386	7.473	9.888	40.884	3.397	0.003
BERTurk[3]	–9.651	3.622	–17.162	–2.139	–2.664	0.014
$R^2 = 0.69$	$F_{(2-22)} = 24.005$	$p < 0.001$	N = 25			

In the analysis detailed in Table 8, only NLP-based similarity models were used. Accordingly, it can be stated that this equation allows for the evaluation of text coherence/cohesion without conducting any additional structural examination or analysis of the text. This analysis explained 69% of the variance in text coherence/cohesion. Furthermore, the model was found to be significant as a result of the regression

analysis ($F[2 – 22] = 24.005$, $p < 0.01$). The regression equation for text coherence/cohesion is presented below:

$$\text{Text Coherence/Cohesion} = (12.318) + (25.386[X_1]) + (–9.651[X_2])$$

$$X_1 = \text{TF-IDF}^4 \qquad X_2 = \text{BERTurk}^3$$

With the completion of the multiple linear regression analyses and the establishment of the cohesion/coherence equations, the regression analyses were concluded. However, in interpreting the results obtained from these equations, specific reference ranges were required. To determine these ranges, the relationship between the participants' mean evaluations of the texts in the sample and their categorical assessments was examined. Subsequently, the coherence/cohesion value of each text was calculated for each regression equation. Based on participants' evaluations, common interpretation ranges were then established for all formulas. These reference ranges are presented in Table 9 below:

Table 9: Reference ranges for the interpretation of results

Coherence Result	Interpretation
0–5.4	Incoherent
5.4–7.1	Partially Coherent (The coherence of the text should be reviewed)
7.1–10	Coherent

With the completion of the analyses described above and the attainment of the findings, methodologies for determining coherence and cohesion that take into account the structural characteristics of agglutinative languages, and specifically Turkish as an agglutinative language, were established. At this stage, the final research question, "Do artificial intelligence models provide valid results in determining the coherence and cohesion of texts?" was addressed, and the performance of AI tools in identifying text coherence and cohesion was examined. The findings of this analysis are presented in Table 10 below:

Table 10: Relationship between AI tools and text coherence

	V	M.	Std. Dev.	1	2	3	4	5	6	7	8	9	10
1	Coherence	6.09	2.54	1									
2	ChatGPT – Coherence	5.96	2.406	0.808**	1								
3	ChatGPT – Cohesion	6.16	1.772	0.864**	0.969**	1							
4	Gemini – Coherence	5.68	3.591	0.817**	0.953**	0.938**	1						
5	Gemini – Cohesion	6	3.416	0.805**	0.948**	0.929**	0.992**	1					
6	Grok – Coherence	5.16	3.338	0.909**	0.919**	0.918**	0.940**	0.917**	1				
7	Grok – Cohesion	5.52	2.694	0.950**	0.890**	0.898**	0.892**	0.878**	0.973**	1			
8	Perplexity – Coherence	6.68	2.577	0.822**	0.945**	0.942**	0.952**	0.970**	0.888**	0.871**	1		
9	Perplexity – Cohesion	6.36	2.691	0.853**	0.935**	0.949**	0.935**	0.952**	0.879**	0.870**	0.979**	1	
10	DeepSeek – Coherence	6.56	3.441	0.833**	0.894**	0.887**	0.912**	0.932**	0.877**	0.871**	0.956**	0.936**	1
11	DeepSeek – Cohesion	6.64	3.067	0.854**	0.890**	0.885**	0.916**	0.931**	0.901**	0.891**	0.944**	0.930**	0.991**

*N = 25; **p < 0.01*

An examination of Table 10 shows that the evaluations of all AI tools from different perspectives exhibited a very high correlation with the assessments of the expert group. In addition, it can be stated that the coherence and cohesion evaluations made with the Grok AI tool measured almost at the same level as the expert judgments.

5 Conclusion

This study aimed to develop methodologies for calculating cohesion and coherence in Turkish by considering the structural characteristics of agglutinative languages. In this context, the coherence and cohesion levels of 25 texts were determined with the assistance of an expert group, and their relationship with the structural features of the texts was subsequently examined. The analyses revealed that the ratio of conjunctions and pronouns (r = 0.694), reference ratio (r = 0.679), demonstrative adjective ratio (r = 0.664), ratio of demonstrative adjectives to pronouns (r = 0.646), conjunction ratio (r = 0.621), cohesive device ratio (r = 0.480), and TTR ratio (r = –0.537) were associated with the coherence and cohesion of texts. By contrast, the preposition ratio was not found to be related to text coherence and cohesion (r = 0.124).

These findings are theoretically consistent with studies that identify cohesion elements such as conjunctions, reference, pronouns, and demonstrative adjectives (Coşkun, 2005; Halliday and Hasan, 1976; He et al., 2025; Lv et al., 2024). However, they cannot be directly compared with previous results, since no empirical research has been conducted to objectively measure the coherence and cohesion of Turkish texts or to examine the factors influencing them. Although relevant studies exist in the international literature, each language has distinct characteristics, making it impossible to directly compare the standards of one language with those of another (Kurudayıoğlu and Karadağ, 2005). In this respect, just as readability formulas lack validity beyond the languages for which they were developed (Çetinkaya, 2010; DuBay, 2004; Kalyoncu and Memiş, 2024; Klare, 1984), statistical analyses of cohesion and coherence will not produce valid results when compared across different languages.

In the research process, beyond examining the structural features of the texts, the relationship between the results obtained from vector-based similarity models and text cohesion and coherence was analyzed. The findings showed that the MiniLM, BERTurk, and TF-IDF models were strongly correlated with cohesion and coherence. However, four different approaches were applied in using these models, and each produced different outcomes. For example, in the regression equations, higher correlations were achieved with BERTurk (r = –0.722) and TF-IDF (r = 0.764) when words were stemmed. This result is meaningful in the context of agglutinative languages, since in Turkish the same word may occur in multiple lemmatized forms depending on context, causing certain NLP models to treat these forms as separate words (Kalyoncu, 2025). Therefore, stemming should be applied prior to using NLP models that are insensitive to the structural characteristics of Turkish and other agglutinative languages, as it leads to more accurate results. However, this approach did not yield valid results for all models. For

instance, with the MiniLM model, the correlation decreased after stemming. Similarly, punctuation cleaning increased the correlation in the TF-IDF model but reduced it in the BERTurk model. Therefore, it is more appropriate to establish a methodology tailored to the characteristics of each model. In this study, the highest correlations were obtained with raw text analysis for MiniLM, stemming-based text analysis for BERTurk, and stemming combined with punctuation cleaning for TF-IDF. By contrast, the FastText model did not perform as reliably as the other models in determining text cohesion and coherence. Based on these findings, it can be concluded that NLP models intended for use in agglutinative languages should be tested under different approaches to validate their effectiveness.

As with the structural features of texts, the results obtained from the NLP models in this study cannot be directly or objectively compared with the existing literature, since these models have not previously been used to detect textual similarity and coherence in Turkish. However, a comparable model, LSA, has been widely applied in measuring cohesion and coherence in English texts. For example, Foltz et al. (1998) introduced LSA as a tool for assessing text coherence, and Graesser et al. (2004) incorporated LSA into their Coh-Metrix tool as an indicator of semantic coherence. Beyond English, studies conducted in other languages have employed models such as FastText and BERT. Research has shown that FastText- and BERT-based metrics strongly correlate with human judgments (Gao et al., 2020; Rei et al., 2020; Zhang et al., 2020). He et al. (2024) examined discourse coherence in patients with psychosis using tools such as FastText and BERT, together with features such as pronoun usage, TTR, and syntactic complexity, and found that these tools had different functions in reflecting discourse coherence. Moreover, scholars have observed that while static similarity models are effective in capturing surface-level coherence, context-sensitive models such as BERT yield more differentiated results. This issue has also been addressed in other studies (Corcoran et al., 2018; Morgan et al., 2021), where discourse coherence was evaluated using NLP-based similarity models. On the other hand, some researchers have argued that existing NLP-based models are insufficient for identifying coherence and have proposed alternatives such as DiscoScore (Zhao et al., 2023). At the same time, Ueda (2021) found that in Japanese, an agglutinative language, the BERT model outperformed traditional methods and was closely related to textual cohesion. In this respect, Ueda's findings align with the present study in the broader context of agglutinative languages.

After examining the relationships between the structural features of the texts, the NLP-based similarity models, and text cohesion and coherence, regression analyses were conducted, resulting in the development of two equations. In the first equation, conjunction and pronoun ratios were incorporated alongside the NLP models, whereas in the second equation only the NLP models were used. The relationship between the results obtained from these equations and the expert evaluations is presented in Table 11 below.

Table 11: Consistency of the formulas with expert evaluations

	V	Mean	Std. Dev.	1	2
1	Coherence	6.099333	2.540152	1	
2	Equation 1	5.347539	2.256449	0.882**	1
3	Equation 2	6.099309	2.094362	0.828**	0.939**

An examination of Table 11 indicates that the formulas developed in this study showed a high correlation with expert evaluations and produced measurements at a comparable level. In this respect, it can be stated that the formulas generated in the study can be effectively used to determine the cohesion and coherence of texts. Moreover, the methodologies proposed here demonstrated stronger consistency with expert judgments than most large language models.

Beyond the variables employed to evaluate text cohesion and coherence, artificial intelligence tools were also assessed. The analyses revealed that all AI tools correlated highly with expert judgments, with cohesion-related scores generally producing more valid results. The performance ranking of these tools, from most to least consistent with expert evaluations, was Grok > ChatGPT > DeepSeek > Perplexity > Gemini. Notably, the measurements obtained with Grok were even more consistent than those produced by the formulas developed in this study ($r = 0.950$). This finding demonstrates that AI tools are highly effective in determining the coherence and cohesion of texts.

It does not seem possible to compare the performance of large language models in measuring text cohesion and coherence with previous studies, since the literature review revealed no research in which large language models were used to assess text coherence. However, the results of this study indicate that large language models achieve a performance level close to that of domain experts in evaluating the cohesion and coherence of texts. In this respect, the use of AI tools for such evaluations appears feasible. Although the methodologies developed in this study performed below the Grok large language model, they nonetheless hold promise for future automated tools specifically designed to assess cohesion and readability.

In light of the findings presented, it can be concluded that the levels of cohesion and coherence in texts are influenced by their structural features, that NLP-based similarity models can be applied to agglutinative languages when appropriate preprocessing steps are performed, that text coherence and cohesion can be measured by combining conjunction and pronoun ratios with both static and contextual NLP models, and that large language models are highly effective in assessing text coherence and cohesion. Accordingly, it is recommended that educators and researchers employ the methodologies proposed in this study together with large language models as an objective approach to measuring the cohesion and coherence of Turkish texts. It is further recommended that preprocessing steps such as stemming and punctuation cleaning be applied prior to the use of NLP models for Turkish and other agglutinative languages, that broader and more comprehensive research be conducted on measuring cohesion and coherence in Turkish and other agglutinative languages, and that the methodologies developed in this study be integrated into Turkish readability formulas.

Strengths and limitations of the study

The originality and strength of this research lie in the fact that no empirical or statistically grounded study had previously been conducted to measure cohesion and coherence in Turkish or to identify the factors influencing them. The methodologies developed in this study therefore represent the first attempt to objectively and automatically measure cohesion and coherence in Turkish. Another strength of the study is that the analyses were conducted in an integrated manner, examining cohesion and coherence through structural features, NLP-based similarity models, and large language models. Nevertheless, the study has certain limitations. It was conducted on a sample of only 25 texts, and some of the participants defined as experts were undergraduate students whose expertise had not yet fully developed. These factors represent weaknesses of the research. Another limitation concerns the use of Google Forms for data collection and the non-randomized presentation of questions during this process, which may have led to order effects. In addition, the methodologies developed in this study are not as comprehensive or advanced as the standards of cohesion and coherence established for an international language such as English.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Data availability statement

The dataset and all analysis scripts used in this study are publicly available on GitHub at the following repository: https://github.com/mrkalyoncu/Coherence_Analysis. The repository includes the complete dataset and the code required to reproduce all analyses and results reported in the article.

Acknowledgements

The corresponding author acknowledges the financial support provided by the Turkish Education Foundation (TEV) through a doctoral scholarship.

6 References

Aminovna, B. D. (2022). Importance of coherence and cohesion in writing. *Eurasian Research Bulletin*, *4*, pp. 83–89.

Azadnia, M., Biria, R., Lotfi, A. R. (2016). A corpus-based study of text cohesion by Coh-Metrix: Contrastive analysis of L1/L2 PhD dissertations. *Modern Journal of Language Teaching Methods*, *6*(2), pp. 265–278.

Bui, P. H., Nguyen, N. Q., Loc, T. N., Nguyen, T. V. (2021). A Cross-linguistic approach to analysing cohesive devices in expository writing by Asian EFL teachers. *The Southeast Asian Journal of English Language Studies*, 27(2), pp. 16–30. https://doi.org/10.17576/3l-2021-2702-02

Büyüköztürk, Ş. (2005). *Sosyal bilimler için veri analizi el kitabı: İstatistik, araştırma deseni, SPSS uygulamaları ve yorum.* Ankara: Pegem Akademi.

Büyüköztürk, Ş., Kılıç Çakmak, E., Akgün, Ö. E., Karadeniz, Ş. ve Demirel, F. (2023). *Eğitimde bilimsel araştırma yöntemleri*. Ankara: Pegem Akademi.

Can, A. (2020). *Bilimsel araştırma sürecinde SPSS ile nicel veri analizi*. Ankara: Pegem Akademi.

Collins, J. L. (1998). *Strategies for struggling writers*. New York: The Guilford Press.

Cook, R.D., Weisberg, S. (1982). *Residuals and influence in regression*. London: Chapman & Hall.

Corcoran, C. M., Carrillo, F., Fernandez-Slezak, D., Bedi, G., Klim, C., Javitt, D. C., Bearden, C. E., Cecchi, G. A. (2018). Prediction of psychosis across protocols and risk cohorts using automated language analysis. *World Psychiatry,* 17, pp. 67–75. https://doi.org/10.1002/wps.20491.

Coşkun, E. (2005). *İlköğretim Öğrencilerinin öyküleyici anlatımlarında bağdaşıklık, tutarlılık ve metin elementleri*. Gazi University. (PhD thesis).

Covington, M. A., McFall, J. D. (2010). Cutting the gordian knot: The moving-average type–token ratio (MATTR). *Journal of Quantitative Linguistics*, 17(2), pp. 94–100. https://doi.org/10.1080/09296171003643098.

Craney, T. A., Surles, J. G. (2002). Model-dependent variance inflation factor cutoff values. *Quality Engineering*, 14(3), pp. 391–403. https://doi.org/10.1081/QEN-120001878.

Creswell, J. W. (2017). *Karma yöntem araştırmalarına giriş* (Trans. ed. M. Sözbilir). Ankara: Pegem Akademi.

Crossley, S. A., Kyle, K., Dascalu, M. (2019). The Tool for the Automatic Analysis of Cohesion 2.0: Integrating semantic similarity and text overlap. *Behavior research methods*, 51(1), pp. 14–27. https://doi.org/10.3758/s13428-018-1142-4.

Crossley, S. A., Kyle, K., McNamara, D. S. (2016). The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior research methods*, 48(4), pp. 1227–1237. https://doi.org/10.3758/s13428-015-0651-7.

Crossley, S. A., McNamara, D. S. (2009). Computationally assessing lexical differences in L1 and L2 writing. *Journal of Second Language Writing*, 18, pp. 119–135. https://doi.org/10.1016/j.jslw.2009.02.002.

Çetinkaya, G. (2010). *Türkçe metinlerin okunabilirlik düzeylerinin tanımlanması ve sınıflandırılması*. Ankara University. (PhD thesis).

Dascalu, M., Trausan-Matu, S., Dessus, P., McNamara, D. S. (2015). Discourse cohesion: A signature of collaboration. In: Blikstein, P., Merceron, A., Siemens, G. (Eds.). *Proceedings of the fifth international conference on learning analytics and knowledge*, pp. 350–354. New York: ACM.

Davoodi, E., Kosseim, L. (2017). On the contribution of discourse structure on text complexity assessment. *arXiv preprint arXiv:1708.05800*.

De Beaugrande, R. A., Dressler, W. U. (1981). *Introduction to text linguistics*. London: Longman.

De Haas, M., Algera, J., Van Tuijl, H., Meulman, J. (2000). Macro and micro goal setting: In search of coherence. *Applied Psychology*, *49*(3), pp. 579–595. https://doi.org/10.1111/1464-0597.00033.

Diep, G. L., Le, T. N. D. (2024). An analysis of coherence and cohesion in English majors' academic essays. *International Journal of Language Instruction*, *3*(3), pp. 1–21. https://doi.org/10.54855/ijli.24331

DuBay, W. H. (2004). *The principles of readability*. Costa Mesa: Impact Information.

Durbin, J., Watson, G. S. (1971). Testing for serial correlation in least squares regression. *Biometrika*, 58(1), pp. 1–19. https://doi.org/10.2307/2334313

Ersöz, F., Ersöz, T. (2019). *Spss ile istatiksel veri analizi*. Ankara: Seçkin.

Follmer, D. J., Sperling, R. A. (2018). Interactions between reader and text: Contributions of cognitive processes, strategy use, and text cohesion to comprehension of expository science text. *Learning and Individual Differences*, 67, pp. 177–187. https://doi.org/10.1016/j.lindif.2018.08.005.

Foltz, P. W., Kintsch, W., Landauer, T. K. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse processes*, 25(3), pp. 285–307. https://doi.org/10.1080/01638539809545029.

Gao, Y., Zhao, W., Eger, S. (2020). SUPERT: Towards new frontiers in unsupervised evaluation metrics for multi-document summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics,* pp. 1347–1354. Online: Association for Computational Linguistics.

Gasparinatou, A., Grigoriadou, M. (2013). Exploring the effect of background knowledge and text cohesion on learning from texts in computer science. *Educational Psychology*, 33(6), pp. 645–670. https://doi.org/10.1080/01443410.2013.790309.

George, D., Mallery, M. (2010). *SPSS for Windows Step by Step: A Simple Guide and Reference, 17.0 update* (10a ed.). Boston: Allyn & Bacon.

Gómez González, M. D. L. Á. (2013). A reappraisal of lexical cohesion in conversational discourse. *Applied Linguistics*, 34(2), pp. 128–150. https://doi.org/10.1093/applin/ams026.

Graesser, A. C., Jeon, M., Yan, Y., Cai, Z. (2007). Discourse cohesion in text and tutorial dialogue. *Information Design Journal*, 15(3), pp. 199–213. https://doi.org/10.1075/idj.15.3.02gra.

Graesser, A. C., McNamara, D. S., Louwerse, M. M. (2003). What do Readers Need to Learn in Order to Process Coherence Relations in Narrative and Expository Text? In: Sweet, A. P. Snow, C. E. (Eds.), *Rethinking Reading Comprehension*, pp. 82–98. New York: The Guilford Press.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior research methods, instruments, & computers*, 36(2), pp. 193–202. https://doi.org/10.3758/BF03195564.

Gujarati, D. N. (2004). *Basic econometrics*. Ohio: The McGraw-Hill Companies.

Güçlü, İ. (2020). *Sosyal bilimlerde nicel veri analizi örneklerle Spss uygulamaları ve yorumlaması*. Ankara: Gazi Kitapevi.

Hall, S., Basran, J., Paterson, K. B., Kowalski, R., Filik, R., Maltby, J. (2014). Individual differences in the effectiveness of text cohesion for science text comprehension. *Learning and Individual differences*, 29, pp. 74–80. https://doi.org/10.1016/j.lindif.2013.10.014.

Halliday, M.A.K., Hasan, R. (1976). *Cohesion in English*. London: Longman Group UK Limited.

He, J., Long, W., Xiong, D. (2025). Evaluating Discourse Cohesion in Pre-trained Language Models. *arXiv preprint arXiv:2503.06137*.

He, R., Palominos, C., Zhang, H., Alonso-Sánchez, M. F., Palaniyappan, L., Hinzen, W. (2024). Navigating the semantic space: Unraveling the structure of meaning in psychosis using different computational language models. *Psychiatry Research*, 333, pp. 115752. https://doi.org/10.1016/j.psychres.2024.115752

Jassim, A. H. (2023). The role of cohesion in text creation. *Language, Discourse & Society*, 11(1), pp. 159–170.

Kaakinen, J. K., Salonen, J., Venäläinen, P., Hyönä, J. (2011). Influence of text cohesion on the persuasive power of expository text. *Scandinavian journal of psychology*, 52(3), pp. 201–208. https://doi.org/10.1111/j.1467-9450.2010.00863.x.

Kalyoncu, M. R. (2025). *Türkçe için yeni bir okunabilirlik formülü*. Ondokuz Mayıs Üniversitesi. (MA thesis).

Kalyoncu, M. R., Memiş, M. (2024). Türkçe için oluşturulmuş okunabilirlik formüllerinin karşılaştırılması ve tutarlılık sorgusu. *Ana Dili Eğitimi Dergisi*, 12(2), pp. 417–436. https://doi.org/10.16916/aded.1434650

Karatay, H. (2010). Bağdaşıklık araçlarını kullanma düzeyi ile tutarlı metin yazma arasındaki ilişki. *Mustafa Kemal Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, 7(13), pp. 373–385. https://izlik.org/JA64TJ63SB

Klare, G. R. (1984). Readability. In: Pearson, P. D. (Ed.) *Handbook of reading research*, pp. 681–744. London: Longman.

Kouti, M. (2022). The development of discourse analysis. *Journal El-Bahith in Human's and Social's Sciences*, 14(1), pp. 523–528.

Kurudayıoğlu, M., Karadağ, Ö. (2005). Kelime hazinesi çalışmaları açısından kelime kavramı üzerine bir değerlendirme. *Gazi Üniversitesi Gazi Eğitim Fakültesi Dergisi*, 25(2), pp. 293–307. https://izlik.org/JA22JC54JJ.

Lismay, L. (2020). Analysis of Coherence and Cohesion on Students' Academic Writing: A case study at the 3rd year students at English education program. *Alsuna*, 3(2), pp. 74–82. https://doi.org/10.31538/alsuna.v3i2.980

Lv, K., Liu, X., Guo, Q., Yan, H., He, C., Qiu, X., Lin, D. (2024). Longwanjuan: Towards systematic measurement for long text quality. *arXiv preprint arXiv:2402.13583*.

McNamara, D. S., Crossley, S. A., McCarthy, P. M. (2010). Linguistic features of writing quality. *Written Communication*, 27(1), pp. 57–86. https://doi.org/10.1177/0741088309351547

McNamara, D. S., Louwerse, M. M., Graesser, A. C. (2002). *Coh-Metrix: Automated cohesion and coherence scores to predict text readability and facilitate comprehension*. Memphis: University of Memphis.

Morgan, S. E., Diederen, K., Vértes, P. E., Ip, S. H. Y., Wang, B., Thompson, B., Demjaha, A., De Micheli, A., Oliver, D., Liakata, M., Fusar-Poli, P., Spencer, T. J., McGuire, P. (2021). Natural language processing

markers in first episode psychosis and people at clinical high-risk. *Transl. Psychiatry* 11, pp. 1–9. https://doi.org/10.1038/s41398-021-01722-y.

O'reilly, T., McNamara, D. S. (2007). Reversing the reverse cohesion effect: Good texts can be better for strategic, high-knowledge readers. *Discourse processes*, 43(2), pp. 121–152.

Ozuru, Y., Dempsey, K., McNamara, D. S. (2009). Prior knowledge, reading skill, and text cohesion in the comprehension of science texts. *Learning and Instruction*, 19(3), pp. 228–242. https://doi.org/10.1016/j.learninstruc.2008.04.003.

Rahma, A., Chelsea, J., Agustina, I. W. (2022). The seven standards of textuality in news texts: A discourse analysis. *STAIRS: English Language Education Journal*, 3(2), pp. 123–138. https://doi.org/10.21009/stairs.3.2.6.

Reed, D. K., Kershaw-Herrera, S. (2016). An examination of text complexity as characterized by readability and cohesion. *The Journal of Experimental Education*, 84(1), pp. 75–97. https://doi.org/10.1080/00220973.2014.963214.

Rei, R., Stewart, C., Farinha, A. C., Lavie, A. (2020). COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2685–2702. Online: Association for Computational Linguistics.

Ringle, C. M., Wende, S., Becker, M. (2015). *SmartPLS 3*. Bönningstedt: SmartPLS.

Sanders, T. J., Maat, H. P. (2006). Cohesion and coherence. In: Brown, K. (Ed.) *Encyclopedia of Language & Linguistics (Second Edition)*, pp. 591–595. Amsterdam: Elsevier.

Thompson, S. (1994). Aspects of cohesion in monologue. *Applied Linguistics*, 15(1), pp. 58–75. https://doi.org/10.1093/applin/15.1.58.

Ueda, N. (2021). BERT-based cohesion analysis of Japanese texts. *Journal of Natural Language Processing*, 28(2), pp. 705–709. https://doi.org/10.5715/jnlp.28.705.

van Dijk, T. A. (1981). Discourse studies and education. *Applied Linguistics*, *2*(1), 1–26. https://doi.org/10.1093/applin/II.1.1

Zhang, T., Kishore, V., Wu, F., Weinberger, K., Artzi, Y. (2020). *Bertscore: Evaluating text generation with BERT*. In 8th International Conference on Learning Representations (ICLR 2020). Addis Ababa: International Conference on Learning Representations.

Zhao, W., Strube, M., Eger, S. (2023). DiscoScore: Evaluating text generation with BERT and discourse coherence. *arXiv preprint arXiv:2201.11176*.