

Master's Thesis for obtaining the academic degree **Master of Arts** in Faculty of Arts and Social Sciences

Exploring Hybrid Linguistic Features for Automatic Readability Assessment in Turkish

Author: Ahmet Yavuz Uluslu

Student ID Number: 21-737-937

Supervisor: Prof. Dr. Gerold Schneider Department of Computational Linguistics Submission date: 01.06.2023

Abstract

We present the first comprehensive study on Automatic Readability Assessment (ARA) for Turkish, an agglutinative language with complex morphological structures. Despite the importance of ARA in various sectors like healthcare, education, and communication, the lack of resources and previous work has left Turkish ARA relatively unexplored. We aim to fill this gap by combining state-of-the-art neural network models with linguistic features at the lexical, morpho-syntactic, syntactic, and discourse levels to create an advanced readability tool. We evaluate the effectiveness of traditional readability formulae compared to modern automated methods and identify key linguistic features that determine the readability of Turkish texts. We also investigate the variance in the readability of popular science magazine articles across different age groups and the performance of state-of-the-art neural and hybrid ARA methods. Lastly, we explore the potential of artificial datasets generated by large language models (LLMs) to supplement training data for Turkish ARA. The findings offer a significant advancement in the field of Turkish readability studies and have important implications for various sectors.

Zusammenfassung

Wir präsentieren die erste umfassende Studie zur automatischen Lesbarkeitsbewertung (ARA) für Türkisch, eine agglutinative Sprache mit komplexen morphologischen Strukturen. Trotz der Bedeutung von ARA in verschiedenen Bereichen wie Gesundheitswesen, Bildung und Kommunikation ist die türkische ARA aufgrund fehlender Ressourcen und früherer Arbeiten noch relativ unerforscht. Wir wollen diese Lücke schließen, indem wir modernste neuronale Netzwerkmodelle mit linguistischen Merkmalen auf lexikalischer, morphosyntaktischer, syntaktischer und Diskursebene kombinieren, um ein fortschrittliches Lesbarkeitswerkzeug zu schaffen. Wir bewerten die Effektivität traditioneller Lesbarkeitsformeln im Vergleich zu modernen automatisierten Methoden und identifizieren linguistische Schlüsselmerkmale, die die Lesbarkeit türkischer Texte bestimmen. Außerdem untersuchen wir die Varianz der Lesbarkeit von populärwissenschaftlichen Zeitschriftenartikeln in verschiedenen Altersgruppen und die Leistung moderner neuronaler und hybrider ARA-Methoden. Schließlich untersuchen wir das Potenzial künstlicher Datensätze, die von großen Sprachmodellen (LLMs) generiert werden, um Trainingsdaten für türkische ARA zu ergänzen. Die Ergebnisse stellen einen bedeutenden Fortschritt auf dem Gebiet der türkischen Lesbarkeitsstudien dar und haben wichtige Auswirkungen auf verschiedene Bereiche.

Acknowledgement

I would like to take this opportunity to thank everyone who supported me with my master's thesis:

I want to express my sincere appreciation to Prof. Dr. Gerold Schneider and Dr. Simon Clematide for introducing me to the exciting field of computational linguistics. I am grateful for their support and collaboration on many interesting projects we have worked on together over the past two years.

Special thanks to Can Yildizli for his mentorship over the years and his unwavering support for my ideas. His guidance has been invaluable in shaping my understanding of the world and research.

I extend my heartfelt thanks to my family and all those who have supported me in completing my master's degree.

Contents

Α	bstract	i
Α	cknowledgement	iv
C	ontents	\mathbf{v}
Li	ist of Figures	vii
Li	ist of Tables	viii
1	Introduction	1
	1.1 Motivation \ldots	. 1
	1.2 Research Questions	2
	1.3 Thesis Structure	3
2	Background	4
	2.1 Early Readability Formulas	4
	2.2 Limitations of Traditional Readability Formulas	7
	2.3 Methods and Techniques in Automatic Readability Assessment	. 8
	2.3.1 Feature Engineering-based Methods	8
	2.3.2 Neural and Hybrid Approaches	10
3	Corpora	12
	3.1 TUBITAK PopSci Magazine Readability Corpus	13
	3.2 LLM-Generated Readability Dataset	14
4	Methodology	18
	4.1 Linguistic Features	18
	4.1.1 Traditional Features (TRAD)	18
	4.1.2 Syntactic Features (SYNX)	19
	4.1.3 Lexico-Semantic Features (LXSM)	21
	4.2 Experiments	24
	4.2.1 Baseline: Non-Neural Models with Linguistic Features	25
	4.2.2 Non-Neural Models	25

	4.2.3 Neural Models	26
	4.2.4 Hybrid Model	27
	4.2.5 Evaluation Metrics	28
	4.2.6 Cross-Validation	29
	4.2.7 In-domain and Cross-Domain Evaluation	29
		01
5	Results	31
	5.1 In-Domain Evaluation Results	31
	5.2 Baseline: Feature and Model Evaluation	32
	5.3 Cross-Domain Evaluation \ldots	33
	5.4 In-Domain + Synthetic Data Evaluation	34
	5.5 Error Analysis	34
	5.6 Model Interpretation	38
	5.6.1 Feature Permutation	38
	5.6.2 Mean Decrease in Impurity (MDI)	39
	5.6.3 Feature Correlation	40
	5.7 Discussion	40
	5.7.1 Linguistic Features	40
	5.7.2 Limitations	41
6	Conclusion	42
	6.1 Future Work	43
Re	eferences	45

List of Figures

1	Different Architectural Approaches in ARA	11
2	Example of a multi-turn synthetic data generation	17
3	Hybrid model architecture	27
4	In-domain evaluation confusion matrix	35
5	In-domain evaluation confusion matrix	36
6	Cross-domain evaluation confusion matrix	36
7	In-domain evaluation confusion matrix	37
8	Cross-domain evaluation confusion matrix	37
9	Feature Importance Ranking by Permutation	38
10	Feature Importance Ranking by MDI	39

List of Tables

1	Descriptive Corpus Statistics	13
2	Readability features across reading levels	13
3	Example sentences for three reading levels	14
4	Candidate article simplification prompt.	15
5	Synthetic Dataset Descriptive Statistics	16
6	List of traditional or surface-based features used for readability as-	
	sessment	19
7	List of phrase and dependency related syntactic features used for	
	readability assessment	19
8	List of tree-depth related syntactic features used for readability as-	
	sessment	20
9	List of part-of-speech related syntactic features used for readability	
	assessment	20
10	Lexical Variation Features	21
11	Type-Token Ratio (TTR) Features	22
12	List of psycholinguistic features used for readability assessment	22
13	List of word familiarity features used for readability assessment	23
14	List of entity density features used for readability assessment	24
15	Hyperparameters used for the non-neural models	26
16	In-domain evaluation results	31
17	Performance comparison of readability models	32
18	Incremental contribution of each feature to the RandomF model	32
19	In-domain vs Cross-domain Evaluation Results	33
20	Training with Synthetic Data Evaluation Results	34
20	Top ten features ranked by their Spearman correlation coefficients	40

1 Introduction

1.1 Motivation

Automatic Readability Assessment (ARA) is an important task in computational linguistics that aims to automatically determine the level of difficulty of understanding a written text, which has implications for various fields, such as healthcare, education, and accessibility Vajjala, 2021). In the healthcare sector, for instance, medical practitioners can use ARA tools to ensure patient information and consent forms are easily understandable Ley and Florio, 1996). In the field of education, teachers and learners alike can benefit from ARA systems to adapt materials to the appropriate language proficiency level Kintsch and Vipond, 2014). Moreover, ARA promotes accessibility by enabling threatomisation of text content to suit diverse reading abilities, enhancing comprehension and engagement for all readers. The appropriate readability of technical reports and other business documents is critical to ensure that the intended audience can fully understand the content and can make informed decisions Bushee et al., 2018). In areas such as cyber-security, readability is particularly important as it can impact response time to risk closures and case materials [Smit et al., 2021].

The task of assessing readability presents challenges, particularly when dealing with large corpora of text. Manual extraction and calculation of linguistic features are time-consuming, expensive, and prone to human errors, leading to subjective assessments [Deutsch et al., 2020]. Recent research in the field has focused on developing automated methods for extracting linguistic predictors and training models for read-ability assessment.

Despite these crucial applications and developments, the ARA task for Turkish, an agglutinative language with complex morphological structures, remains relatively unexplored. This is mainly due to the lack of resources and previous work in this area. Recent readability efforts in Turkish have largely been confined to traditional readability formulae, such as Flesch-Kincaid Kincaid et al., 1975 adaptations [Ateşman, 1997; Bezirci and Yilmaz, 2010; Çetinkaya, 2010]. These measures

typically rely on superficial text features such as sentence length and word length. Several previous studies have demonstrated the unreliability of these formulas with practical examples Feng et al., 2010, 2009. The integration of complex morphological, syntactic, semantic, and discourse features in modern ARA approaches offers the possibility of significantly improving the current readability studies in Turkish.

1.2 Research Questions

We present the first comprehensive ARA study for Turkish. Our study combines traditional raw text features with lexical, morpho-syntactic, syntactic, and discourse information to create an advanced readability tool. We demonstrate the effectiveness of our tool on a new readability corpus we created from popular science magazine articles, published for different age groups and educational levels. Our research aims to answer the following research questions:

- 1. How well do state-of-the-art neural and hybrid approaches perform in the task of readability assessment for Turkish texts? How do these methods compare to traditional formulae and each other?
- 2. What are the key linguistic features, at various levels (morphological, syntactic, semantic, and discourse), that determine the readability of Turkish texts and how can these features be effectively incorporated into an ARA system?
- 3. How does the readability of popular science magazine articles in Turkish vary across different age groups? Can this variance be accurately captured and modeled using automated ARA methods?
- 4. Can artificial datasets generated by large language models (LLMs) effectively supplement the dataset used for Turkish readability assessment? How does the performance of models trained on this supplemented dataset compare to their performance on the original dataset?

1.3 Thesis Structure

This thesis is organised into six chapters. Chapter 2 provides an overview of the history, applications, and limitations of traditional readability formulae. A detailed analysis of state-of-the-art automatic readability assessment models is then provided, highlighting how they differ from traditional approaches in that they use advanced computational techniques to account for deeper linguistic structures and nuances not considered in earlier methods. Chapter 3 introduces the Turkish readability corpus we created based on popular science articles for young audiences, and explains how we used the power of LLMs to synthetically generate a complimentary parallel corpus for our training and testing purposes. Our exploration of hybrid linguistic features and neural models specifically tailored for Turkish is presented in Chapter 4. In Chapter 5, we undertake a rigorous evaluation of different approaches and present the results of our investigation. Finally, Chapter 6 concludes the thesis by summarising our findings and offering a thorough analysis of the implications of our research.

2 Background

In this chapter, we explore the landscape of automatic readability assessment. Section 2.1 provides a historical perspective, examining early readability formulae and their applications. Section 2.2 addresses the limitations inherent to these traditional formulas, underscoring their tendency to oversimplify text complexity. The advent of ARA, a response to these limitations, is introduced in Section 2.3. It details the methodologies and techniques employed in ARA, specifically the integration of machine learning and natural language processing (NLP). In Section 2.4, we analyse the implications of these methodologies and challenges specifically for Turkish readability, recognising the need to consider diverse linguistic features and state-of-the-art models in automatic readability assessment.

2.1 Early Readability Formulas

The research of quantifying text readability, or the ease with which a text can be read, has a history spanning over a century DuBay, 2007. Initial research was centered on the creation of lists of difficult words and readability formulas such as Flesch Reading Ease Flesch, 1948, Dale-Chall readability formula Dale and Chall, 1948, Gunning FOG Index Gunning, 1969 and SMOG Mc Laughlin, 1969. These formulas are essentially simple weighted linear functions that utilize easily measurable variables such as word and sentence length, as well as the proportion of complex words within a text. This foundational work marked the onset of a systematic approach to assessing text readability. They were swiftly adopted in a wide range of fields such as medicine, education, and communication, emphasising their broad applicability.

The Flesch Reading Ease formula Flesch, 1948, represented as Equation 2.1, quickly gained popularity due to its simplicity and efficiency in assessing text readability. It was used to assess the readability of technical manuals in the U.S. Army to ensure that these important documents are easily understood by military personnel. This widespread adoption led to the development of a variant known as the Flesch-

Kincaid Grade Level formula Kincaid et al. [1975] which provides a readability score corresponding to U.S. school grade levels.

$$FRE = 206.835 - 1.015 \left(\frac{total_words}{total_sentences} \right) - 84.6 \left(\frac{total_syllables}{total_words} \right)$$
(2.1)

Equation 1: Fleisch Reading Ease formula.

Initially developed for the English language, the Flesch Reading Ease formula required recalibration for its application to Turkish, a task undertaken by Atesman [Ateşman, 1997]. This recalibration was necessary due to the unique linguistic characteristics and structure of Turkish compared to English. However, a significant obstacle in its adoption was Atesman's failure to disclose the statistical variables used in the recalibration process. This gap was later addressed in the work of Cetinkaya-Yilmaz [Cetinkaya] 2010, which also assigned appropriate grade levels, thus facilitating its practical use in the Turkish educational context. Despite these initial challenges, both recalibrated formulas Atesman and Cetinkaya-Yilmaz have become the most popular readability assessment tools in Turkish. In Equation 2.2 and Equation 2.3] X_1 represents the average word length in syllables, and X_2 represents the average sentence length in words.

$$Atesman = 198.825 - 40.175 \times x_1 - 2.610 \times x_2 \tag{2.2}$$

$$Cetinkaya = 118.823 - 25.987 \times x_1 - 0.971 \times x_2 \tag{2.3}$$

Not long after the adaptation, Bezirci-Yilmaz Bezirci and Yilmaz, 2010 introduced an important refinement, akin to the approach taken in the SMOG formula. They posited that the impact of polysyllabic words on text complexity is distinct from the total number of syllables present in the text. Accordingly, they include the counts of polysyllabic words (those with 3-, 4-, and 5+ syllables) as separate features in their readability formula given in Equation 2.4

$$BY = \sqrt{AWS \times ((S3 \times 0.84) + (S4 \times 1.5) + (S5 \times 3.5) + (S6 \times 26.25))}$$
(2.4)

Equation 2: Bezirci-Yilmaz Readability formula.

In the formula above, BY represents the readability score. The variables used in

the formula are defined as follows: AWS is the average number of words per sentence in the text, S3 represents the count of polysyllabic words with 3 syllables, S4 represents the count of polysyllabic words with 4 syllables, S5 represents the count of polysyllabic words with 5 syllables, and S6 represents the count of polysyllabic words with more than 6 syllables.

In parallel, Sönmez 2003 encountered inconsistencies when applying the Gunning FOG Index to Turkish texts. This observation led to the development of a new mathematical formula, now known as the Sonmez formula, specifically designed for Turkish texts.

$$Sonmez = \left(\frac{meaning_rate}{AWS}\right) \times difficulty_rate$$
(2.5)

Equation 4: Sonmez Readability formula.

In Equation 2.5 the "meaning rate" refers to the number of foreign words, phrases, terms, concepts, figurative expressions, similes, formulas, symbols, etc., present in the text divided by the total number of sentences in the text. It represents the density of unfamiliar or abstract elements in the text. The "difficulty rate" represents the number of foreign words, terms, formulas, figurative expressions, symbols, and concepts divided by the total number of words in the text, giving an indication of the complexity or difficulty level of the text. The limitations are primarily due to its subjective nature when identifying complex words and concepts, which stands in contrast to other formulas that utilise more objective criteria, such as syllable counts, to evaluate text complexity.

Readability formulae have found practical applications in several areas of readability studies in Turkish, particularly in the fields of medicine and education. In medical contexts, these formulae are used to gauge the complexity of Patient Education Materials (PEMs) and other health-related texts, such as consent forms. These tools help to ensure that crucial health information is accessible and understandable to all, regardless of their educational background. For instance, researchers have used the Flesch-Kincaid and Atesman Readability formulae to assess the readability of anesthesia consent forms in Turkish hospitals, which led to valuable insights into how these documents could be optimized for better comprehension [Boztas et al.] [2017]; Boztaş et al.] [2014]. In the realm of education, readability studies have been employed to evaluate the complexity of textbooks, thereby ensuring that these crucial learning materials are appropriate for the targeted student age group. For example, research has been conducted to determine the readability levels of Turkish tales

in middle-school textbooks, providing insights that could potentially enhance the quality of education by aligning learning materials with students' comprehension abilities [Turkben, 2019; Tekşan et al., 2020; Guven, 2014]. While traditional read-ability formulas have significantly contributed to the field of readability assessment, they are not without their limitations. They often rely heavily on surface-level text features, such as word and sentence length, and fail to account for deeper linguistic and cognitive factors that influence readability [Collins-Thompson, 2014].

2.2 Limitations of Traditional Readability Formulas

Readability formulae have inherent limitations that can affect their accuracy and applicability. One of the most significant challenges is the necessity for languagespecific re-calibration. Given the unique phonetic attributes, sentence formation patterns, and mean syllable length inherent in each language, each language requires its own calibrated readability formula. The validity of studies employing readability formulae calibrated for the English language to evaluate texts in other languages remains questionable. In practice, applying an English-calibrated formula to Turkish texts may result in an overestimation of readability levels. Indeed, most studies that have used this approach have reported inflated readability requirements Akgül, 2019; Akgül, 2022. This discrepancy can be attributed to the differences in language structure and usage, emphasizing the need for language-specific calibration. Furthermore, the evolution of language over time may necessitate periodic recalibration of these formulas Lee and Lee, 2023. As language trends evolve and new words and phrases become commonplace, readability formulas must adapt to remain accurate and relevant. The research indicates that traditional readability measures display unreliable performance when applied to non-traditional document types, such as web pages Petersen and Ostendorf, 2009.

Traditional readability formulas, despite their extensive use, have been criticised for their lack of wide linguistic coverage [Feng et al., 2009, 2010]. These formulas predominantly focus on superficial text features, largely ignoring other linguistic aspects that significantly contribute to text readability. Factors such as syntactic and semantic complexity, discourse structure, and other linguistic branches recognised by [Collins-Thompson, 2014] which are integral to comprehending a text, remain largely unaccounted for in these traditional models. This narrow linguistic focus can lead to inaccuracies in readability assessment, especially when applied to languages or texts with diverse linguistic structures. These scores are relative measures of readability that should be interpreted in the context of the text's overall features and the target audience's reading ability. They are not absolute measures and treating them as such can result in a misunderstanding of the text's actual readability.

Practitioner errors in applying readability formulas often stem from methodological shortcomings and misinterpretations [Wang et al., 2013]. The requirement of traditional measures for considerable text sample sizes introduces another impediment, even though the theoretical minimum size for a text sample has yet to be conclusively established. For instance, a common methodological error is the inappropriate sampling of text. Some studies might only consider a limited section of a text, such as the first 100 words, leading to skewed results, especially in scientific texts where complexity often increases later in the document. Similarly, the selective assessment of text sections that do not accurately mirror the overall complexity of the text, like focusing solely on the introduction or conclusion, can misrepresent the readability level.

2.3 Methods and Techniques in Automatic Readability Assessment

ARA is predominantly characterised as a machine learning classification problem within the NLP discipline. However, alternative approaches have also been explored, including modelling it as a regression problem [Vajjala and Meurers] 2014 and a ranking problem [Tanaka-Ishii et al.] 2010; Lee and Vajjala] 2022. Current ARA methodologies largely depend on a gold standard training corpus, annotated with labels that designate reading level categories or numbers that illustrate a graded scale [Vajjala] 2021]. This annotated corpus forms the foundational basis upon which subsequent steps in the ARA pipeline are built. ARA pipeline involves the construction of the readability model. This critical step encompasses both feature extraction/text representation and the training of an ARA model. Over the past two decades, research dedicated to developing readability models has primarily focused on language models and feature engineering-based machine learning approaches.

2.3.1 Feature Engineering-based Methods

In his work, Collins-Thompson 2014 offers a comprehensive analysis of text features that affect readability, dividing them into four distinct linguistic branches. Lexico-semantic features encompass aspects such as rare, unfamiliar, or ambiguous words, which can significantly influence comprehension. Morphological features focus on

the presence of rare or complex morphological particles that could pose challenges to readers. Syntactic features refer to the grammatical structure of the text, which affects sentence processing and thus the overall complexity of a text [Hale, 2016]. Finally, discourse features concern the structure and coherence of larger text units, integral to the overall readability of the text.

Lexico-semantic features often involve the relative frequency of a word in everyday use or its presence/absence in a reference word list. Given the established connection between reading comprehension and vocabulary knowledge, which is in turn related to the frequency of word usage, it is reasonable to conclude that vocabulary frequency serves as a reliable predictor of text readability or reading difficulty [Chen and Meurers, 2018]. In readability models, these features can be used individually, representing distinct lexical properties, or in an aggregated form, providing collective estimates of lexical difficulty. An example of such an aggregated feature is the type-token ratio (TTR) shown in [2.3.1], a measure of lexical richness that reflects the range and diversity of vocabulary used in a text. Additionally, statistical language models can serve as a source of lexical features providing the relative probability of encountering any given vocabulary word in a text.

$$textTTR = \frac{unique_words(types)}{words(tokens)}$$
(2.6)

Equation 3: Type-Token Ratio (TTR) formula.

Morphological features especially in languages rich in inflectional and derivational morphology such as Turkish, can play a significant role in assessing readability. These features pertain to the choice of word suffixes and prefixes that convey meaning. For instance, Hancke et al. [2012] demonstrated the efficacy of incorporating additional morphological features in readability classification for German.

Syntactic features constitute another important dimension in automated readability assessment. It is known to be associated with longer sentence processing times [Gibson, 1998]. Rather than merely relying on traditional sentence length, modern approaches harness a more diverse set of features to assess a text's syntactic complexity. Syntactic complexity features are subsequently derived from natural language parsers, providing a richer understanding of text readability.

Discourse features represent the higher-level structures in a text that stem from the dependencies and relationships among their elements [Pitler and Nenkova, 2008]. Cohesion refers to the interpretation of one element in a text depending on another,

while coherence reflects the logical ordering of arguments and ideas as well as the systematic organisational structure of the text. These discourse-level structures often determine the overall flow of the text, playing a significant role in shaping its readability.

The aforementioned features serve as input variables for machine learning models. Notably, feature engineering-based methods often employ feature selection strategies to isolate a subset of features from this larger set. This selection process is instrumental in improving the model's predictive performance by prioritizing those features that contribute most significantly to the readability of the text. The resultant feature vectors, each representing a text, are then ingested by the model for training, thus allowing the model to learn the complex function that maps these features to a readability score. François and Miltsakaki [2012] conducted a study comparing the performance of classic and non-classic readability features, using two predictive models, namely linear regression and Support Vector Machines (SVMs). They reported that the exclusion of non-classic predictors detrimentally impacted prediction performance, and the optimal best performance was achieved by incorporating both classic and non-classic features.

2.3.2 Neural and Hybrid Approaches

In recent years, there has been an emerging trend toward utilizing neural models for ARA. These models have demonstrated the capacity to implicitly capture the previously mentioned features without the need for manual feature extraction Jawahar et al., 2019. Martinc et al. 2021 and Imperial 2021 experimented with contextual embeddings of BERT Devlin et al., 2018 for the readability assessment task, achieving par or better results than feature-based approaches. Meng et al. [2020] introduced ReadNet, a hierarchical self-attention-based Transformer model, where hierarchical structure allows for the encapsulation of information at various text levels (word, sentence, paragraph), thereby capturing varying complexities that contribute to overall readability. They demonstrated that transformer outputs serve as effective features on non-parallel readability assessment datasets. In these datasets, documents exhibiting different levels of readability tend to have considerable semantic disparities. However, both studies omitted cross-domain evaluation, leading to uncertainty about the extent to which language models rely on topic and genre information, as opposed to readability. Other studies have further explored various strategies to integrate linguistic features with transformer models, promoting a fusion of traditional and neural approaches Lee et al., 2021; Deutsch et al. 2020. The state-of-the-art results are currently being achieved by hybrid models that ensemble linguistic features with transformer-based models, highlighting the combined strength of traditional and modern approaches. Figure 1 illustrates three distinct ARA architectures, with a notable emphasis on the hybrid approach.



Figure 1: Different Architectural Approaches in ARA

3 Corpora

Training data for ARA can originate from a diverse range of sources. These sources can be largely categorized into two groups: expert annotated and non-expert annotated. The former refers to resources such as textbooks or graded readers, which are meticulously crafted by trained authors with a specific target audience in mind. Such expert-annotated materials represent the most prevalent form of training data employed in ARA research. Non-expert annotated sources include web content or data gathered through crowdsourcing experiments and user studies. Most widely used readability corpora include One Stop English (OSE) [Vajjala and Lučić, 2018], the WeeBit corpus Vajjala and Meurers, 2012 and the Newsela corpus Xu et al., 2015]. The OneStopEnglish (OSE) dataset was created by collecting articles from an English language learning resources website. The dataset contains 189 articles, each of which was manually rewritten into two simpler versions. It includes three difficulty classes: elementary, intermediate, and advanced. The OSE dataset is considered parallel, as the semantic distinctions among the three classes are minimal. The focus of the rewriting process was primarily on simplifying syntax and substituting complex words with simpler alternatives. The WeeBit dataset comprises 6,388 English articles that are categorized into five grade levels covering ages 7-16 years. The three lower levels were extracted from the WeeklyReader newspaper and the two higher levels were from the BBC-Bitesize corpus. Therefore it is a non-parallel dataset with articles from different levels. While the majority of these benchmark datasets and corpora are predominantly available in English, there is a growing interest in the development of readability corpora in other languages. In the context of low-resource languages, limited access to digital text resources necessitates reliance on conventional learning materials, such as classroom materials and textbooks. There are currently no existing readability corpora available for Turkish.

3.1 TUBITAK PopSci Magazine Readability Corpus

Our corpus was constructed using popular science articles from TUBITAK Popular Science Magazines¹ spanning the period 2007 to 2022. The articles are openly published and generously made available for non-commercial redistribution and research purposes. We selected 2250 articles from three magazines, each catering to readers of different age groups. These magazines include Merakli Minik (for ages 0-6), Bilim Cocuk (for ages 7+), and Bilim ve Teknik (for ages 15+). Accordingly, we consider the articles from these magazines as elementary, intermediate, and advanced level reading material. Our corpus is non-parallel and encompasses a diverse range of topics, including instructions for laboratory experiments and brief articles about recent scientific discoveries. This characteristic is similar to that of the WeeBit corpus, which also includes articles from various topics and resources. Given that the articles in our corpus are written by experts and specifically tailored for distinct age groups, it can be appropriately regarded as an 'expert-annotated' corpus. We used a off-the-shelf pdf-to-text converter to extract the relevant article text and manually corrected the articles to ensure the conversion accuracy of Turkish characters and the layout integrity. Table 1 displays descriptive statistics for the finalized corpus.

Reading Level	Avg. Words	Std. Dev	Nr. of Articles
Elementary	120.95	67.35	750
Intermediate	154.99	93.57	750
Advanced	327.08	187.54	750

Table 1: Descriptive Corpus Statistics

As expected, the advanced texts display a greater average length compared to the elementary texts. However, the high standard deviation values for each level indicate that other factors beyond text length may have a significant impact on determining the reading level of a given text.

Feature	ELEMENTARY	INTERMEDIATE	ADVANCED
Atesman	66.06	59.73	42.32
Cetinkaya	39.31	36.62	29.81
TTR	0.65	0.71	0.76

Table 2: Readability features across reading levels

¹https://yayinlar.tubitakx.gov.tr/

We also performed preliminary analysis on the three reading levels of the corpus using traditional formulae and showed the results in Table 2 presenting readability metrics Atesman, Cetinkaya-Uzun, and Type-Token Ratio (TTR). As expected, the Atesman and Cetinkaya readability scores decrease from one level to the next indicating that texts become more complex at higher reading levels. In contrast, the TTR score increases suggesting that texts become more diverse and less repetitive at higher reading levels. It should also be noted that the readability levels of the elementary-level articles in both formulas were not suitable for the intended age group and that the magazine's disclaimer states that certain articles may require the assistance of an adult or parent. Table 3 presents examples of articles representing each of the three reading levels.

Reading Level	Example	
Elementary Burası bir doğa koruma merkezi. Burada annesi ve babası olmayan t ları var. Merkezde çalışanlardan biri özel bir giysi giyip koluna bir tu geçirmiş		
Intermediate	Robotlar, insanların yaptığı işleri, onların yerine yapan karmaşık makinelerdir. Bu işleri yapmak için programlanırlar. Otomatik olarak ya da uzaktan kumanda edilerek belirli komutları yerine getirirler	
Advanced	Pek çok canlıda manyetik algının varlığı bilimsel olarak biliniyor. Bakteri, salyangoz, kurbağa ve ıstakoz gibi canlılar Dünya'nın manyetik alanını algılıyor, göçmen kuşlar ve deniz kaplumbağaları yönlerini bu sayede buluyor, köpekler eğitildiklerinde saklanmış çubuk mıknatısın yerini gösterebiliyor	

Table 3: Example sentences for three reading levels

3.2 LLM-Generated Readability Dataset

The use of LLMs such as GPT-3 Brown et al., 2020 for data annotation was shown to drastically reduce costs associated with manual human labeling. As demonstrated by Wang et al. 2021a, when used as a weak annotator in conjunction with fewshot prompting and active learning, it delivered results that are about twenty times cheaper than human labeling. In further demonstrating the potential of LLMs, recent research has shown the ability of models such as ChatGPT to outperform traditional crowd-workers in a variety of annotation tasks such as relevance, stance, topics, and frames detection Gilardi et al., 2023. The zero-shot accuracy of Chat-GPT exceeded that of crowd-workers in four out of the five of the tasks evaluated. Apart from text annotation, Wang et al. 2021b presented the capabilities of fewshot prompting on a large pre-trained language model to synthesize high-quality data for training purposes. The results from their approach were found to be better or at least comparable to those from strong baseline models that were trained on humanlabeled data. When both synthetic and human-labeled data were combined, they achieved new state-of-the-art results on the SuperGLUE benchmark. Michail et al. 2023 explored the implications of using automatically generated examples with ChatGPT in low-resource tasks. They worked on the task of predicting the level of intimacy in tweets, a multilingual setting for which limited data was available. Notably, even with a small number of synthetic data generated, a slight increase was observed for low-resource languages and overall accuracy. The successful application of large language models (LLMs) such as GPT3.5 and ChatGPT in sentence simplification tasks, as demonstrated by Feng et al. [2023] and Jeblick et al. [2022], motivated us to consider a similar approach in our work. Notably, they achieve performance comparable to human reference simplifications and outperform current state-of-the-art sentence simplification methods, particularly in multilingual tasks. Recent investigations by Jiao et al. 2023 and Bang et al. 2023 indicate that Chat-GPT competes favorably with established commercial translation systems, thereby demonstrating its capacity for generating multilingual content.

	Zero-Shot Multi-Turn Article Generation Prompt		
	This is an advanced level article for age 15+ from a popular science magazine for		
children. Keep the meaning same but make it simpler for age group $\{7-15\}$			
	{intermediate} level.		
	Complex: {Advanced Article}		
P1	Simple: {Outputs}		
	Keep the meaning same but make it simpler for age group $\{0\mathchar`-7\}$ and $\{\mbox{elementary}\}$		
	level.		
	Simple: {Outputs}		

Table 4: Candidate article simplification prompt.

Our research adopts a similar prompt to that of Feng et al. [2023] within a zeroshot and multi-turn setting. The concept of zero-shot learning refers to the model's capability to tackle tasks without having encountered any explicit examples during training. In order to avoid complexities associated with compiling a collection of unrelated documents for each reading level such as topic and genre bias [Vajjala, 2021], our approach concentrates on constructing a set of triplet graded readers to serve as a parallel dataset. Rather than generating new articles from scratch, we opted to select advanced-level articles and simplify them into two levels. We synthesise 50 simplified articles with ChatGPT (GPT 3.5) corresponding to each

Level	Avg. Words	Std. Dev	Atesman	Cetinkaya	TTR	Nr. of Articles
Elementary	143.75	59.97	60.43	36.43	0.73	50
Intermediate	219.67	118.80	52.01	32.74	0.75	50
Advanced	347.73	181.70	40.01	27.90	0.78	50

distinct readability level. The simplification prompt used in the study is shown in Table 5.

Table 5: Synthetic Dataset Descriptive Statistics.

To ensure the quality of the synthetic dataset, manual checks were conducted to confirm the grammar, structural integrity, and correctness of the information within the generated articles. The same descriptive statistics applied to our readability corpus were subsequently generated for our newly created synthetic dataset, with results displayed in Figure 2. The length of the generated articles aligns with the standard deviation of the initial readability corpus, maintaining a similar trend. This observation strengthens the resemblance of our synthetic dataset to the original corpus in terms of text length and its relation to readability levels. In line with other findings, we observe that the Atesman and Cetinkaya readability scores decrease with increasing reading levels, suggesting greater text complexity. Conversely, the TTR score increases, suggesting more diverse and less repetitive language at higher reading levels. The consistent trends observed in the descriptive statistics of the synthetic dataset and the original corpus underline the potential of using the generated synthetic data as a complementary resource to the original corpus in readability studies.

Hodel: Default (GPT-3.5)

This is an advanced level article for age 15+ from a popular science magazine for children. Keep the meaning same, but make it simpler for age group 7-15, intermediate level.

Son iki yüzyılda gerçekleştirilen kültür tarihi, özellikle de bilim tarihi çalışmaları, Ortaçağ İslam dünyasında 750 1300 yılları arasını kapsayan sürede yüksek bir uygarlık yaratıldığını, Nefis gibi birçok bilim ve düşün insanının yetiştiğini ortaya koymuştur. Ulaşılan bu yüksek uygarlık 12. yüzyıldan başlayarak ilerleme hızını kaybetme ye başlamış, 14. yüzyıla gelindiğinde ise tamamen dur muştur. Bir zamanlar bilimin öncülüğünü yapan İslam dün yası, artık bilim ve felsefenin rüzgarının dindiği, yeni bilim ve düşün insanlarının yetişmediği verimsiz, çorak bir toprağa dönüşmüştür. Başta Osmanlılar olmak üzere zaman zaman bu kötü gidişi durdurmak ve entelektüel kültürün öncülüğünü yeniden ele almak için çaba gösterilmişse de, bu çabalar kötü gidişi durdurmaya yetmemiş, İslam dün yası uygarlık yarışından tamamen kopmuştur. Bugün de bireysel veya bölgesel bazı gelişmeler kaydedilse de, 8. ve 14. yüzyıllar arasında görülen görkemli uygarlık ateşi bir daha hiç yakılamamıştır. İslam dünyasın da duraklamanın başladığı sıralarda ise, yani 12. yüzyıl dan itibaren Batı yoğun ve sistemli bir çeviri hareketi başlatarak İslam dünyasının seçkin bilim ve düşün yapıtlarını kendi kültürüne kazandırmış ve 14. yüzyıldan itibaren entelektüel kültürün önderliğini üstlenerek bugüne kadar getirmiştir. Bu kısa betimlemenin apaçık sonucu şudur: Bilim takdir edildiği topraklarda yeşermekte, takdir edilmediğinde ise o topraklardan göç etmektedir.

G

Son iki yüzyılda yapılan araştırmalar, Ortaçağ İslam dünyasında 750-1300 yılları arasında büyük bir uygarlık olduğunu gösterdi. Bu dönemde birçok bilim insanı ve düşünür yetişti. Ancak, bu yüksek uygarlık 12. yüzyıldan itibaren yavaşladı ve 14. yüzyılda tamamen durdu. İslam dünyası eskiden bilim ve felsefenin öncüsüyken, artık bilim ve düşünceye önem verilmeyen, verimsiz bir toplum haline geldi. Bazı çabalara rağmen, Osmanlılar gibi, bu kötü gidişat durdurulamadı ve İslam dünyası uygarlık yarışından koparıldı. Günümüzde bazı bireysel veya bölgesel gelişmeler olsa da, 8. ve 14. yüzyıllar arasında görülen büyük uygarlık ateşi tekrar yakılamadı. İslam dünyasında duraklama dönemi başladığında, yani 12. yüzyıldan itibaren Batı, İslam dünyasının bilim ve düşünce eserlerini kendi kültürüne çevirerek ve 14. yüzyıldan itibaren de entelektüel kültürün liderliğini üstlenerek günümüze kadar getirdi. Bu durum bize şunu gösteriyor: Bilim değer verilen yerlerde gelişir, değer verilmediğinde ise o yerlerden ayrılır.

Keep the meaning same, but make it simpler for age group 0-7, elementary level.

凸 凸 구 ◀»

Bir zamanlar İslam dünyasında çok zeki insanlar yaşardı. Onlar bilim ve düşünceyle ilgilenirlerdi. Ama sonra bu uygarlık yavaşladı ve durdu. Artık bilim ve düşünceye önem vermiyorlardı. Bu nedenle İslam dünyası geri kaldı. Bazı yerlerde küçük gelişmeler oldu, ama eskisi gibi büyük başarılar elde edilemedi. Bu sırada Batı dünyası İslam dünyasının bilim ve düşünce eserlerini kendi kültürüne çevirdi ve ilerledi. Bu nedenle bugün Batı daha ileride. Önemli olan bilime değer vermektir. Eğer bilime değer verirsen, büyük gelişmeler kaydedebilirsin.

Figure 2: Example of a multi-turn synthetic data generation

4 Methodology

4.1 Linguistic Features

In this study, we explore five subgroups of linguistic features from our Turkish readability corpus: traditional or surface-based features, syntactic features, lexico-semantic features, morphological features, and discourse features. We employ spaCy v3.4.0 [Honnibal et al., 2020] with the pre-trained tr_core_news_trf model [Turkish NLP Suite, 2022] for the majority of general tasks, including sentence/entity recognition, POS tagging, and dependency parsing. The model was trained on the Universal Dependencies BOUN Treebank [Marşan et al., 2022] and Turkish Wiki NER dataset [Altinok, 2022]. We use the Stanford Stanza parser version 1.5.0 [Qi et al., 2020] for constituency parsing.

4.1.1 Traditional Features (TRAD)

Traditional or surface-based features are commonly used to predict the readability of Turkish texts, and we also adopt them as a baseline for our study. Specifically, we extract 7 traditional features shown in Table 6 including Turkish adaptations of well-known readability formulas such as Atesman and Cetinkaya-Uzun, as well as average values of words and syllables. As noted by Bezirci and Yilmaz 2010 in their evaluation of the Turkish readability formulae, the impact of the number of polysyllabic words on text complexity is different from that of the total number of syllables present in the text. Therefore, we also included the counts of polysyllabic words (3-, 4-, and 5+ syllables) as separate features in our analysis. We intentionally avoid extracting aggregate features such as total word or total sentence count to predict readability, as these tend to be inherently dependent on text length. Although the complexity of a text may depend on these features, our corpus size could constrain the applicability of the results derived from such metrics. Hence, our focus is primarily on average values and ratios that offer a more balanced view of text complexity, irrespective of text length.

Feature name	Description
w₋ps	Average number of words per sentence
syll₋pw	Average number of syllables per word
syll₋ps	Average number of syllables per sentence
poly4_ps	Average number of words with 4 or more syllables per sentence
poly5₋ps	Average number of words with 5 or more syllables per sentence
poly6_ps	Average number of words with 6 or more syllables per sentence
atesman	Readability score according to Atesman's formula
cetinkaya	Readability score according to Cetinkaya-Uzun's formula

Table 6: List of traditional or surface-based features used for readability assessment

4.1.2 Syntactic Features (SYNX)

Syntactic properties have a significant impact on the overall complexity of a given text, which serves as an important indicator of readability. We extract an array of syntactic features that capture various dimensions of sentence structure.

Phrasal and dependency type features: Reading abilities are related to the ratios involving clauses in a text Lu, 2010. We extract features based on noun and verb phrases available at sentence and article levels. We integrate features based on the unconditional probabilities of their dependency-based equivalents Dell'Orletta et al., 2011. These encompass various types of syntactic dependencies, including subject, direct object, and modifier, among others. This is demonstrated in Table 7, which presents the list of phrasal and dependency-type features used in our readability assessment model.

Feature Name	Description
a₋np	Average number of noun phrases per word and sentence
a₋vp	Average number of verb phrases per word and sentence
a₋pp	Average number of prepositional phrases per word and sentence
u_prob_sub_dep	Unconditional probability of subject dependencies
u_prob_do_dep	Unconditional probability of direct object dependencies
u_prob_mod_dep	Unconditional probability of modifier dependencies



Parse tree depth features: The depth and structure of dependency trees in a text can reflect the level of sentence complexity. Following this principle, we extract features shown in Table 8 representing both the average and maximum depths of the constituency and dependency tree structures present in the text [Dell'Orletta et al.], 2011].

Feature Name	Description
a_dep_tree_depth	Average dependency tree depth per sentence
max_dep_tree_depth	Maximum dependency tree depth per article
a_const_tree_depth	Average constituency tree depth per sentence
max_const_tree_depth	Maximum constituency tree depth per article

Table 8: List of phrase and dependency related syntactic features used for readability assessment

Part-of-Speech features: Part-of-speech (POS) tags provide essential information about the syntactic function of words in sentences. Adapting the work of Tonelli et al. 2012 and Lee et al. 2021, we include features based on universal POS tag counts. Such features offer insights into the distribution and usage of different word categories, adding another layer of syntactic information. The overview of POS features was demonstrated in Table **Q**.

Feature Name	Description
a₋adj	Average number of adjectives per word and sentence
a₋adv	Average number of adverbs per word and sentence
a₋adp	Average number of adpositions per word and sentence
a_coord_conj	Average number of coordinating conjunctions per word and sentence
a₋subord₋conj	Average number of subordinating conjunctions per word and sentence
a₋numeral	Average number of numerals per word and sentence
a₋noun	Average number of nouns per word and sentence
a_verb	Average number of verbs per word and sentence
a_prop_noun	Average number of proper nouns per word and sentence
a₋pronoun	Average number of pronouns per word and sentence
a_func_word	Average number of function words per word and sentence
a_cont_word	Average number of content words per word and sentence

Table 9: List of part-of-speech related syntactic features used for readability assessment

4.1.3 Lexico-Semantic Features (LXSM)

Lexico-semantic features are a set of linguistic attributes that can reveal the complexity of a text's vocabulary. These features can be used to identify specific words or phrases that may pose difficulty or unfamiliarity to readers [Collins-Thompson], 2014].

Lexical Variation features: Secondary language acquisition research has found a correlation between the diversity of words within the same Part-Of-Speech (POS) category and the lexical richness of a text Vajjala and Meurers [2012]. We report noun, verb, adjective, and adverb variations, which represent the proportion of the respective category's words to the total. For instance, the feature often termed "verb variation" in related literature is calculated by dividing the total number of unique verbs by the total count of verbs in a text. As outlined in Equations 6 and 7, we extracted simple (VV) and root lexical variations (RVV) shown in Table 10.

$$VV = \frac{unique_verbs}{total_verbs}$$
(4.1)

Equation 6: Simple Verb Variation

$$RVV = \frac{\sqrt{unique_verbs}}{\sqrt{total_verbs}}$$
(4.2)

Equation	7:	Root	Verb	Variation
----------	----	------	------	-----------

Variation Measure	Description
Noun Variation	Simple and root noun variations
Verb Variation	Simple and root verb variations
Adjective Variation	Simple and root adjective variations
Adverb Variation	Simple and root adverb variations

Table 10: Lexical Variation Features

Type Token Ratio (TTR) features: TTR is a commonly used metric to quantify lexical richness and has been widely employed in readability assessment studies. We compute five distinct variations of TTR from [Vajjala and Meurers, 2012] shown in Table 11] The standard TTR variations of a text sample are susceptible to the text length, which can introduce bias in the readability assessment. To address this limitation, we also consider the Moving-Average Type–Token Ratio (MATTR) [Covington and McFall, 2010]. The MATTR mitigates the length-dependency issue by calculating the TTR score within a moving window across the text.

Measure	Formula	Description	
Root TTR	T/\sqrt{N}	Type count divided by the square root of token count	
Corrected TTR	T/(2N)	Type count divided by twice the token count	
Bilogarithmic TTR	$\log T / \log N$	Log of type count divided by log of token count	
Uber Index	$\log^2 T \log N/T$	Square of log type count multiplied by log token count	
		divided by type count	
MTLD	MTLD	Measure of Textual Lexical Diversity (TTR 0.72)	
MAATR	MAATR	Moving-Average Type-Token Ratio	

Table 11: Type-Token Ratio (TTR) Measures

Psycholinguistic features: We adopted word frequencies obtained from the Turkish psycholinguistic database created by Acar et al. [2016]. This resource was built from transcriptions of children's speech and corpora of children's literature, thus containing words commonly acquired during early development. It also includes words typically acquired during adulthood from a standard corpus. We calculated the average word and sentence frequency for both early and late-acquired words in the article. Table [12] presents the psycholinguistic features used in our readability assessment, including average log10 values [Brysbaert and New, 2009] per word and sentence for the early and late acquired word corpora.

Feature Name	Description
e_freq_pw	Average log10 value per word in early corpus
e_freq_ps	Average log10 values per sentence in early corpus
l_freq_pw	Average log10 value per word in late corpus
l_freq_ps	Average log10 value per sentence in late corpus

Table 12: List of psycholinguistic features used for readability assessment

Word Familiarity features: Familiarity with specific words can greatly affect readability. Based on prior work on Italian Dell'Orletta et al., 2011 and French François and Fairon, 2012 readability studies, we assessed the vocabulary composition of the articles using a reference list of 1700 basic words essential for achieving elementary reading proficiency in Turkish. This list, a combination of the first 1200 words taught to children aged 0-6 [Keklik, 2010] and a set of essential words from

an open-access textbook¹ for learning Turkish, provides a benchmark for vocabulary familiarity. We calculated the percentage of unique words (types) in the text based on this reference list, performed on a lemma basis. These features are summarized in Table 13.

Feature Name	Description
avg_unk_w_s	Average number of unknown words per sentence from the list
avg_unk_w_a	Average number of unknown words per article from the list
prop_known_w	Proportion of known words based on the reference list

Table 13: List of word familiarity features used for readability assessment

Morphological features: Morphological complexity plays a significant role in readability assessment, particularly in languages that are morphologically richer than English, such as German and Basque. This has been demonstrated in the works of Hancke et al. [2012] and Gonzalez-Dios et al. [2014]. Moreover, morphological complexity has been leveraged in cross-lingual readability assessment frameworks, proving its applicability across languages with varying morphological structures [Weiss et al.] [2021]. However, these studies have not explored agglutinative languages like Turkish and Hungarian. In our study, we integrate the Morphological complexity Index (MCI), as proposed by [Brezina and Pallotti, 2019]. The MCI captures the variability of morphological exponents of specific parts-of-speech within a text by comparing word forms with their stems. We calculate MCI features for verbs, nouns, and adjectives, considering different sample sizes and sampling techniques with and without repetition.

The equation for the MCI is as follows:

$$MCI = \frac{WSV + BSD}{2} - 1 \tag{4.3}$$

where:

- WSV (Within-subset variety) measures the diversity of inflectional types within each subset of words. For instance, if a subset of verbs includes "run" (tr: koş), "runs" (tr: koşacak), and "running" (tr: koşuyor), the within-subset variety would be 3 because there are three different inflectional types.
- BSD (Between-subset diversity) measures the diversity of inflectional types across all subsets. For example, if one subset of verbs includes "run" (tr: koş),

¹https://www.turkishtextbook.com/most-common-words/

"runs" (tr: koşacak), and "running" (tr: koşuyor), and another subset includes "jump" (tr: zıpla), "jumps" (tr: zıplayacak), and "jumping" (tr: zıplıyor), the between-subset diversity would be 6 because there are six different inflectional types across all subsets.

Discourse features: The final group of features we examine are Entity Density Features (EnDF). The presence and frequency of entities within a text can significantly impact the cognitive load required for comprehension. Entities often introduce new conceptual information, thereby increasing the burden on the reader's working memory. This relationship between entities and readability was previously studied by Feng et al. [2009, 2010], whose entity-related features we incorporate into our analysis. The entity density features used in our study are listed in Table 14.

Feature Name	Description
a_entity_person	Average number of PERSON entities per word and sentence
a_entity_quantity	Average number of QUANTITY entities per word and sentence
a_entity_org	Average number of ORG entities per word and sentence
$a_{-}entity_{-}ordinal$	Average number of ORDINAL entities per word and sentence
a_entity_norp	Average number of NORP entities per word and sentence
a_entity_loc	Average number of LOC entities per word and sentence
a_entity_event	Average number of EVENT entities per word and sentence
a_entity_gpe	Average number of GPE entities per word and sentence
a_entity_date	Average number of DATE entities per word and sentence
a₋entities	Average number of entities per word and sentence

Table 14: List of entity density features used for readability assessment

4.2 Experiments

We detail the set of experiments that we designed to examine the effectiveness of various automated readability assessment (ARA) approaches on Turkish texts. The purpose of these experiments is to not only establish baseline performance metrics for different linguistic features but also to elucidate the potential benefits of more complex neural and hybrid models in the context of Turkish readability assessment. We experiment with four different setups: trad-baseline (non-neural model with shallow features), modern-baseline (non-neural model with linguistic features), neural (pre trained transformer models), and hybrid (modern-baseline + neural). We assess these approaches across three different settings. The first is an in-domain setting, using 10-fold cross-validation solely on the TUBITAK readability corpus. The second setting, termed in-domain synthetic, trains on a combination of the readability corpus and the synthetic parallel dataset, but evaluates only on the readability corpus test set. Finally, the cross-domain setting trains on the readability corpus, but tests on the synthetic parallel dataset.

4.2.1 Baseline: Non-Neural Models with Linguistic Features

Given the lack of available baselines for the readability task in Turkish, our first objective is to establish a baseline for the readability task. This baseline is designed to be on par with traditional readability formulas and is reliant on shallow linguistic features such as sentence and word lengths. Non-neural machine learning algorithms, well-documented for their performance in diverse tasks, form the foundation of our model. By establishing this baseline, we are effectively creating a benchmark that allows for meaningful comparison between the traditional readability formulas, which are the only available methods in readability assessment for Turkish. Following the establishment of this baseline, we expand our feature set and include a more diverse set of linguistic feature groups. Our focus here is twofold: not only are we interested in the performance of individual features, but we also aim to identify the best-performing combinations when these features are assembled into linguistic groups.

4.2.2 Non-Neural Models

For our non-neural readability assessment models, we have chosen to employ a range of traditional machine learning algorithms that have proven effective for a variety of classification tasks. Specifically, our baseline models consist of Logistic Regression (LR), Support Vector Machines (SVM), Random Forest (RF), and XGBoost [Chen et al., 2015]. These algorithms have been extensively used in readability studies, and they have yielded promising results. However, it is worth noting that over the last half-decade, the landscape of readability assessment has been reshaped with the advent of neural models. Despite their relatively recent introduction, these models are rapidly gaining traction due to their potential to discern complex patterns and structures within the data.

To ensure the optimal performance of our models, we perform a two-step hyperparameter tuning process. First, we carry out a randomized search to explore a broad range of hyperparameter values. Following this, we apply a grid search to identify the optimal combination of hyperparameter values within this narrowed range. Grid search is an exhaustive search method that scans through the manually specified subset of the hyperparameter space to find the best hyperparameters for model training. In Table 15, we provide an overview of the best-performing hyperparameters for each non-neural model during the hyperparameter tuning process.

Model	Hyperparameters	Values
Logistic Regression (LR)	С	0.1, 0.5, 1 , 2, 5
	Penalty	L2, L1
	Solver	newton-cg, lbfgs, liblinear, saga
Support Vector Machines (SVM)	С	0.1, 0.5, 1 , 2, 5
	Gamma	scale, auto
	Kernel	rbf, sigmoid, poly
Random Forest (RF)	n_estimators	5, 7 , 10, 15, 25
	max_depth	5, 10, None
	max_features	sqrt, log2, auto
XGBoost	eta	0.01, 0.05, 0.03 , 0.1
	gamma	0, 1 , 3, 5
	max_depth	3, 4, 6 , 10

Table 15: Hyperparameters used for the non-neural models

4.2.3 Neural Models

We extend from the established usage of BERT [Deutsch et al.] 2020; Martinc et al.] 2021 and RoBERTa [Lee et al.] 2021 in readability assessment and opt for the BERTurk model² from Huggingface transformers library [Wolf et al.] 2020] for our analysis. By addressing certain limitations in BERT's training process, such as dynamic masking and larger batch sizes, it has been shown to outperform BERT in several benchmarks [Liu et al.] 2019. We tested multiple learning rates and batch sizes to ascertain the optimal configuration for our task. Specifically, we examined the learning rates of [1e-5, 2e-5, 3e-5, 1e-4] and the batch sizes of [8, 16, 32]. Our final model used AdamW optimizer [Loshchilov and Hutter], 2017], linear scheduler with 10% warmup steps, batch size of 8, and learning rate of 3e-5. The sequence lengths of our input documents were all set to 512 tokens. We fine-tune our model for five epochs.

²https://huggingface.co/dbmdz/bert-base-turkish-uncased

4.2.4 Hybrid Model

In our study, we experiment with a hybrid model approach that aims to leverage the strengths of both neural and non-neural models in an ensemble learning strategy. The premise behind the hybrid model is based on the observation that while neural models such as BERT have demonstrated robust performance across diverse tasks, they could still benefit from incorporating handcrafted linguistic features, which have been key components in traditional non-neural models [Deutsch et al.] 2020].

Our hybrid model takes a straightforward approach: it combines the soft label predictions generated by the neural model with handcrafted features. This composite is then used as input to a non-neural model. The non-neural model, which we refer to as the secondary predictor, learns from two sources: the predictions or outputs of the neural model (the initial predictor), and the handcrafted features. This architecture allows us to bolster the performance of the secondary predictor by supplementing what the initial predictor might miss. Specifically, the inclusion of handcrafted features serves to reinforce the model's understanding of the data, leading to more robust performance. This approach is similar to the methodologies adopted by Imperial 2021 and Lee et al. 2021, who demonstrated the effectiveness of combining neural and non-neural models for improved performance for the readability task.



Figure 3: Hybrid model architecture

4.2.5 Evaluation Metrics

In our work, we employ several evaluation metrics to quantify the performance of our models. These metrics, including accuracy, recall, precision, and the F1 score, are widely used in the fields of machine learning and information retrieval.

• Accuracy (Acc): Accuracy is the ratio of correctly predicted instances to the total instances. It quantifies the frequency of correct predictions made by our model on the readability level. Accuracy is computed using the formula:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$
(4.4)

• Recall (Rec): Also known as sensitivity or the true positive rate, recall measures the proportion of actual positive instances that are correctly identified as such. It reflects our model's ability to find all relevant cases within a dataset. The recall is calculated as follows:

$$Recall = \frac{TP}{TP + FN} \tag{4.5}$$

• **Precision (Prec):** Precision calculates the proportion of positive identifications that are actually correct. It reflects our model's ability to return only relevant instances. Precision is given by the formula:

$$Precision = \frac{TP}{TP + FP} \tag{4.6}$$

• F1 Score (F1): The F1 score is a metric that considers both precision and recall to compute the trade-off between the two. It is particularly useful in situations where class distributions are imbalanced. The F1 score is defined as the harmonic mean of precision and recall:

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$
(4.7)

In this set of metrics, TP represents the number of true positives, TN represents the number of true negatives, FP represents the number of false positives, and FN represents the number of false negatives.

4.2.6 Cross-Validation

To assess the performance of our model, we apply a cross-validation scheme, a common validation technique used for assessing how the results will generalize to an independent dataset. The primary advantage of cross-validation lies in its ability to leverage the entire dataset for both training and testing purposes. This is particularly important when dealing with small datasets like ours where splitting the data into separate training and test sets may lead to insufficient data for effective model training.

We chose to apply a ten-fold cross-validation (10FCV) scheme. In this setting, the dataset was split into ten equally sized groups, each containing roughly 450 documents. The cross-validation process was conducted as follows: in each of the five iterations, four of the groups were used for model training, and the remaining group was used for testing. The performance of our model was evaluated in terms of accuracy, as well as F1, Recall, and Precision for the classification task. The final reported results for each metric are the average values obtained from the five separate test sets. This approach helps to provide a more robust estimation of the model's performance.

4.2.7 In-domain and Cross-Domain Evaluation

Previous studies have investigated the impact of cross-domain data in the task of ARA. For instance, it was observed that ranking models trained on the Newsela dataset suffered a decrease in performance when tested on the OneStopEnglish and Vikidia datasets Lee and Vajjala, 2022. Recently, Lim et al. 2022 attempted to address this challenge by leveraging two comparable datasets to conduct a crossdomain evaluation on hybrid models. Their study specifically focused on assessing the contribution of linguistic features in the readability grade prediction task. In our study, we carry out two distinct sets of experiments to better understand the influences of domain factors and the potential of parallel data in readability assessment. The first experimental setup focuses on cross-domain evaluation. Here, we train our models on the readability corpus, a non-parallel dataset with a broad range of texts, but conduct our evaluation on the synthetically generated parallel dataset. This approach enables us to test the model's ability to distinguish different readability levels of the same article, a scenario that is particularly important for neural models due to their content-focused nature. Furthermore, we evaluate the hybrid approach in this setup to investigate how traditional and neural features might complement each other for this task.

Our second experimental setup centers around in-domain evaluation. In this case, we train our models on a mixed dataset comprising both the readability corpus and the synthetic parallel dataset and then evaluate them on the readability corpus alone. This setup is designed to test if the parallel dataset, with its abundant examples of the same topic written at different readability levels, can serve as a valuable training resource that aids the model in focusing more on readability aspects beyond just the topic. As in the first setup, we also evaluate all three approaches (non-neural, neural, and hybrid) under this experimental design. These distinct experimental setups provide valuable insights into the robustness of our models against cross-domain variability, their ability to capture readability features regardless of content variations, and the potential role of parallel data in enhancing readability assessment.

5 Results

5.1 In-Domain Evaluation Results

The results from our in-domain evaluation, as detailed in Table 18, illustrate a clear hierarchy of performance amongst our tested models. Our Hybrid model, which fuses the strengths of both traditional and neural methodologies, outshines all other models, securing the highest values for accuracy, precision, recall, and F1 score. Following the hybrid model, the neural model performs the best. The neural model (BERT) demonstrates an enhanced ability to capture nuanced characteristics of text readability, exhibiting superior performance to the baseline models without any handcrafted linguistic features. The modern baseline, incorporating five different linguistic subgroups, achieves superior performance compared to the traditional baseline. This highlights the advantage of leveraging an extended set of linguistic features over merely relying on surface-level features typical of traditional readability formulae.

Model	Acc(%)	Prec(%)	Rec (%)	F1 (%)
trad-baseline (shallow)	65.7	67.5	66.8	66.7
modern-baseline (linguistic features)	85.3	85.3	85.1	85.1
neural (BERT)	92.8	93.1	92.6	92.8
hybrid (modern + neural)	96.1	96.1	95.6	95.8

Table 16: In-domain evaluation results

These results suggest that the improved granularity and extent of linguistic features, combined with the ability of neural models to identify complex patterns, significantly contribute to improving the performance of readability assessment. However, prior research has identified potential limitations and biases associated with neural models, particularly when utilised with small, non-parallel corpora - conditions that are present in our current experimental setup. To further investigate these potential constraints and assess the broader applicability of our models, we undertake a subsequent cross-domain evaluation.

5.2 Baseline: Feature and Model Evaluation

The process of feature and model selection for the baseline models was carried out with a careful analysis of the results obtained from different combinations. This section discusses the comparison and performance of traditional and modern baselines to illustrate the process of arriving at the best-performing model.

Model	Acc (%)	Rec (%)	Prec (%)	F1 (%)
SVM	78.1	78.1	79.0	77.6
RandomF	85.3	85.3	85.1	85.1
LogR	83.7	83.6	83.7	83.5
XGBoost	84.1	84.0	84.0	83.7

Table 17: Performance comparison of readability models

Through rigorous evaluation, four distinct models, namely Support Vector Machines (SVM), Random Forest (RandomF), Logistic Regression (LogR), and XGBoost, were assessed using the combination of five different linguistic groups: traditional (TRAD), lexico-semantic (LXSM), syntactic (SYNX), morphological (MORPH), and discourse (DISCO) features. Table 17 provides a comparative view of these models' performance when trained using the full combination. Among the four models evaluated, the Random Forest model delivered the highest performance with 85.3%. This suggests that the Random Forest model, paired with the total combination of linguistic groups, is able to capture the complexity of the readability assessment task more efficiently.

Features Added	Acc (%)
TRAD	65.7
TRAD + LXSM	76.4
TRAD + LXSM + SYN	82.5
TRAD + LXSM + SYN + MORPH	83.6
TRAD + LXSM + SYN + MORPH + DISCO	85.3

Table 18: Incremental contribution of each feature to the RandomF model

Importantly, all of the linguistic groups used provide orthogonal or distinct information. Table 18 effectively illustrates this phenomenon, demonstrating how each uniquely contributing linguistic group incrementally improves the accuracy of the Random Forest model. Their combined strength ultimately achieves the highest overall accuracy score.

5.3 Cross-Domain Evaluation

As anticipated, the performance of the neural models experienced a considerable decline under cross-domain settings. We interpret this decrease in performance as an indication of the diminished capacity of these models for semantic differentiation when evaluating parallel datasets. In this specific context, the models face difficulties in leveraging content information to distinguish between various reading levels, which can be attributed to the inherent characteristics of parallel datasets. Contrary to neural models, the baseline models, both the traditional and modern baselines, demonstrated a certain degree of robustness. Even though their performance also saw a reduction, the extent was considerably less severe. This can be attributed to the extent of linguistic features, which do not predominantly rely on content information. These models primarily utilise a collection of engineered features, which gives them a certain degree of robustness against changes in the content, preserving their performance in cross-domain applications. The hybrid model, which combines the strengths of both linguistic features and neural modeling, retained a higher portion of its performance. This result showcases the potential of handcrafted linguistic features in contributing to the robustness and efficiency of readability assessment models, even in cross-domain settings.

	In-domain		Cross-	domain
Models	Accuracy (%)	F1-score (%)	Accuracy (%)	F1-score (%)
trad-baseline	65.7	66.7	62.1↓	62.0↓
modern-baseline	85.3	85.1	76.8↓	76.6↓
neural	92.8	93.0	70.5↓	70.7↓
hybrid	96.1	95.8	79.5↓	79.2↓

Table 19: In-domain vs Cross-domain Evaluation Results

5.4 In-Domain + Synthetic Data Evaluation

We explore the performance of our models when trained on a combination of indomain and synthetic data. As we incorporate synthetic data into the training process, an upward shift in the performance of all models is observed. While the improvements are slight, they are consistent across all model architectures, pointing to the potential value of the synthetic data in boosting the effectiveness of our models. This prompts a more detailed error analysis, which is vital to determine whether the synthetic data serves to alleviate content bias, and subsequently refines the models' capacity to focus on readability features.

	Readability Corpus		+Synthetic Data	
Models	Accuracy (%)	F1-score (%)	Accuracy (%)	F1-score (%)
trad-baseline	65.7	66.7	66.1 ↑	67.0↑
modern-baseline	85.3	85.1	85.6 ↑	85.4↑
neural	92.8	93.0	93.1 ↑	93.2 ↑
hybrid	96.1	95.8	96.3 ↑	96.0 ↑

Table 20: Training with Synthetic Data Evaluation Results

5.5 Error Analysis

Based on our analysis of the confusion matrix of **modern-baseline** for in-domain evaluation, our findings indicate that the model's ability to accurately assess readability levels improves with every feature subset, particularly in distinguishing advanced level texts from the rest. This observation can be attributed to the fact that elementary (age 0-6) and intermediate texts (age 7+) exhibit a comparable level of lexico-semantic and syntactic complexity, making it more challenging to differentiate between them accurately. This can be visualized in the confusion matrix figure shown in Figure 4, where we can observe the model's performance in classifying texts across different proficiency levels. Additionally, it is important to note that the age groups defined in our corpus encompass an open-ended age range, allowing for variability within each category. Furthermore, we acknowledge that easy texts in our corpus may exhibit varying levels of complexity, as some articles are intentionally written to be read with parental assistance.



Figure 4: In-domain evaluation confusion matrix (modern-baseline)

We observed that our hybrid model significantly improved the differentiation between easy and intermediate readability classes. However, our findings suggest that this improvement is primarily attributed to the neural model's ability to discern topic differentiation and content-based cues rather than the inherent readability features. This hypothesis was substantiated when we evaluated the neural model cross-domain, where the content remained constant across readability levels. Under these conditions, the neural model's proficiency to differentiate between easy and intermediate classes, and also between intermediate and advanced classes, significantly decreased. Consequently, the overall model accuracy dropped to 70.5%. Therefore, while our neural model presents certain advantages, it also underscores the challenges of differentiating readability levels based solely on content features, accentuating the need for further investigation into the role of readability features in model performance. Figure 5 and Figure 6 visually represent the diminished performance of the neural model in cross-domain evaluations, with particular difficulty discerning between adjacent readability classes intermediate and advanced, as well as easy and intermediate. Figures 7 and 8 demonstrate the resilience of the hybrid models to a similar effect, wherein the models maintain a more effective differentiation between adjacent classes.



Figure 5: In-domain evaluation confusion matrix (neural)



Confusion Matrix for Neural Model (Lower Accuracy)

Figure 6: Cross-domain evaluation confusion matrix (neural)



Figure 7: In-domain evaluation confusion matrix (hybrid)



Figure 8: Cross-domain evaluation confusion matrix (hybrid)

5.6 Model Interpretation

In order to gain insights into the significance of individual linguistic features within our best-performing model, the RF model, we utilised two well-established model interpretation techniques specifically designed for Random Forest models: Feature Permutation and Mean Decrease in Impurity (MDI).



Feature Importance (Permutation)

Figure 9: Feature Importance Ranking by Permutation

5.6.1 Feature Permutation

Feature Permutation is a model interpretation method that assesses the importance of a feature by permuting its values and observing the resulting decrease in model accuracy [Breiman, 2001]. This permutation disrupts the feature's original predictive information while preserving its distribution. If a feature is influential in the model's predictions, the permutation will lead to a substantial reduction in accuracy, indicating its importance. Figure 9 visualizes the feature importance as assessed by feature permutation. The x-axis represents the decrease in model accuracy (expressed as a percentage) after the permutation of each feature, while the y-axis lists the features. A higher decrease in accuracy corresponds to a more important feature.



Figure 10: Feature Importance Ranking by MDI

5.6.2 Mean Decrease in Impurity (MDI)

Mean Decrease in Impurity (MDI) is another approach to evaluate feature importance in tree-based models. It calculates the total reduction in a chosen criterion (such as Gini impurity or entropy) achieved by a particular feature Breiman, 2002). The importance of a feature is computed as the normalized total reduction in the criterion, commonly referred to as Gini Importance. Figure 10 presents the MDIbased feature importance evaluation. The x-axis indicates the Mean Decrease in Impurity, and the y-axis lists the features. Higher MDI values denote more crucial features.

5.6.3 Feature Correlation

We also considered model-independent analysis to gain additional perspective into the importance of features with respect to readability levels. Spearman correlation analysis, a non-parametric measure of rank correlation, provides insights into the relationship between each individual feature and readability levels. Table 20 presents the ten features with the highest Spearman correlation coefficients highlighting the significance of readability assessment.

Linguistic Set	Feature	Spearman Correlation
TRAD	Sentence Length Mean	0.487
TRAD	Polysyllable Count	0.467
LXSM	Child Corpus Proportion	0.433
SYNX	Mean Tree Depth	0.419
LXSM	Lexical Verb Variation	0.403
LXSM	Early Frequency PW	0.385
LXSM	Corrected TTR Score	0.352
LXSM	Lexical Density	0.321
LXSM	Lexical Noun Variation	0.297
SYNX	Noun Phrase Per Word	0.278

Table 20: Top ten features ranked by their Spearman correlation coefficients

5.7 Discussion

5.7.1 Linguistic Features

The analysis of feature importance consistently highlights the significant role of traditional features in assessing readability. Particularly, average sentence length and polysyllable counts emerge as powerful features that can effectively capture readability levels. These findings align with previous research, where it has been shown that even compared to more complex feature extraction methods, a simple measure like sentence length can indirectly capture multiple linguistic aspects of readability. Furthermore, our analysis demonstrates that lexico-semantic features play a prominent role in determining readability. This is also evident from the performance improvement observed when including LXSM linguistic feature set in the modernbaseline method. It indicates that while traditional features are indeed valuable, incorporating fine-grained information at the semantic and lexical level can lead to an even better understanding of overall readability. Additionally, the consistent presence of the syntactic feature "mean tree depth" further supports the relationship between sentence length and syntactic complexity. The correlation between mean tree depth and mean sentence length suggests that the structural complexity captured by syntactic features aligns with the overall complexity of sentences.

5.7.2 Limitations

Despite the contributions we make in this study, several limitations should be acknowledged. Firstly, our feature extraction process required separate tools for constituency parsing and dependency parsing. These tools have their preprocessing steps, resulting in computationally expensive operations. Furthermore, while both parsers are available, their accuracy rates and training data are limited. As the sentence complexity increases, noise may be introduced by these parsers. While it may not directly affect aggregated features such as lexical diversity, outliers may be present in fine-grained features such as the maximum statistical values in the article. This aspect should be further investigated and considered in future research. The morphological analyzer used in our study also has its limitations and suffers from the aforementioned issue. We only considered one morphological feature (Morphological Complexity Index), and the sampling window and rate for this algorithm have not been extensively explored. Although our corpus size is sufficient, as the minimum text length is already over a hundred words, the implications of different sampling approaches warrant further investigation.

While previous studies have shown ChatGPT to be a suitable simplification annotator, integrating human-in-the-loop or human annotators into the process may yield even better results. Our study focused on a specific corpus, and there is a lack of benchmark studies in other linguistic registers. It is crucial to understand the implications of our findings and establish benchmarks across different registers, such as textbooks, literature, and health documents, to ensure the applicability of the results in diverse contexts. Furthermore, although we made efforts to highlight certain biases in the model, the applicability of the results to different corpora remains unexplored. It is essential to investigate the generalizability of the proposed approaches and assess their effectiveness in various linguistic domains to gain a more comprehensive understanding of readability assessment.

6 Conclusion

In this thesis, we have made several contributions to the field of readability assessment for the Turkish language. We introduced a new readability corpus specifically designed for popular science magazine articles targeting young audiences, providing a valuable resource for future research in Turkish readability assessment. Furthermore, we leveraged large language models to generate a synthetic parallel dataset based on our readability corpus, enabling us to expand the training data for our models.

By exploring the effectiveness of linguistic features at different levels, we have demonstrated their superiority over traditional readability formulae and shallow-level features. Our findings emphasize the importance of incorporating comprehensive linguistic features in Turkish readability assessment, as they provide more nuanced insights into the complexity of Turkish texts. Additionally, we investigated the potential of hybrid models that combine linguistic features with state-of-the-art transformer models. This hybrid approach proved to be highly effective, achieving improved results compared to both traditional baseline models and standalone neural models. By leveraging the strengths of both linguistic features and neural modeling, we achieved a more comprehensive and accurate assessment of readability.

Furthermore, our cross-domain evaluation of the synthetic parallel dataset revealed the presence of content or topic bias inherent in neural models for the readability task. This bias affects the models' ability to differentiate between reading levels when the content of the article remains the same. However, our hybrid models demonstrated a certain degree of robustness, mitigating the impact of content bias and showcasing the potential of combining linguistic features with neural models to address this challenge.

Lastly, we explored the impact of synthetic data on model performance. Our findings suggest that the inclusion of synthetic data can introduce a slight performance improvement, particularly in scenarios where the model struggles to differentiate the difficulty of texts with similar topics. This highlights the value of synthetic data as a supplementary resource for training readability assessment models. In conclusion, this thesis has significantly advanced the field of readability assessment for the Turkish language. We have introduced a new readability corpus, developed advanced readability models, and explored the effectiveness of linguistic features and hybrid approaches. Our findings shed light on the strengths and limitations of different methodologies, setting up the research for future research in Turkish readability assessment. The insights gained from this study have implications for educational materials, language learning resources, and automated text-processing applications in Turkish, ultimately contributing to a more accessible environment.

6.1 Future Work

We intend to work on several important areas that can further advance the field of Turkish readability assessment. First, expanding the dataset to include the entire collection of magazines from the same resource would provide a more comprehensive and diverse set of texts for evaluation and generalization of readability models. Furthermore, exploring potential resources for creating parallel corpora for the Turkish language is crucial to improve the robustness and effectiveness of readability assessment models, leveraging cross-lingual approaches.

Considering individuals with special needs, such as vowel dyslexia and speech impairment, is an important direction to make readability assessment more inclusive. Adapting readability models to accommodate the specific challenges faced by these individuals would contribute to creating more accessible text-processing applications. Additionally, efforts should be made to enhance the accessibility and usability of existing psycholinguistic data for Turkish, along with developing tools to facilitate its utilization. This will not only benefit readability assessment but also enable further research in areas such as estimating the age-of-acquisition, which is currently lacking in available datasets.

To capture the unique morphological complexity of the Turkish language, additional approaches beyond the morphological features utilized in this study should be explored. Investigating new methods or features specifically tailored to Turkish morphology would provide more accurate assessments of readability. Moreover, comparing the performance of the Turkish model to cross-lingual readability approaches that are gaining popularity would shed light on the model's effectiveness across different languages and uncover potential transfer learning opportunities.

Incorporating coherence chain features in readability models is crucial for capturing text coherence. However, due to the lack of discourse-related corpora, some coherence chain features were not calculable in this study. Therefore, collecting or creating discourse-related corpora for Turkish would be a valuable to enhance the models' ability to evaluate coherence. Additionally, based on the identified important features, future research could focus on developing a new readability formula specifically tailored to the Turkish language. This formula would provide a more comprehensive and accurate measure of readability.

References

- E. A. Acar, D. Zeyrek, M. Kurfalı, and C. Bozsahin. A turkish database for psycholinguistic studies based on frequency, age of acquisition, and imageability. In *Proceedings of the Tenth International Conference on Language Resources* and Evaluation (LREC'16), pages 3600–3606, 2016.
- Y. Akgül. The accessibility, usability, quality and readability of turkish state and local government websites an exploratory study. Int. J. Electron. Gov. Res., 15 (1):62–81, jan 2019. ISSN 1548-3886. doi: 10.4018/IJEGR.2019010105. URL https://doi.org/10.4018/IJEGR.2019010105.
- Y. Akgül. Evaluating the performance of websites from a public value, usability, and readability perspectives: a review of turkish national government websites. Universal Access in the Information Society, pages 1–16, 2022.
- D. Altinok. turkish-nlp-suite/ner-datasets: Newpaper ner set, Oct. 2022. URL https://doi.org/10.5281/zenodo.7265312.
- E. Ateşman. Türkçede okunabilirliğin ölçülmesi. Dil Dergisi, 58(71-74), 1997.
- Y. Bang, S. Cahyawijaya, N. Lee, W. Dai, D. Su, B. Wilie, H. Lovenia, Z. Ji, T. Yu, W. Chung, et al. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. arXiv preprint arXiv:2302.04023, 2023.
- B. Bezirci and A. E. Yilmaz. Metinlerin okunabilirliğinin ölçülmesi üzerine bir yazılım kütüphanesi ve türkçe için yeni bir okunabilirlik ölçütü. Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik Dergisi, 12(3):49–62, 2010.
- N. Boztaş, Ş. Özbilgin, E. Öçmen, G. Altuntaş, S. Özkardeşler, V. Hancı, and A. Günerli. Evaluating the readibility of informed consent forms available before anaesthesia: a comparative study. *Turkish journal of anaesthesiology and reanimation*, 42(3):140, 2014.

- N. Boztas, D. Omur, S. Ozbilgin, G. Altuntas, E. Piskin, S. Ozkardesler, and V. Hanci. Readability of internet-sourced patient education material related to "labour analgesia". *Medicine*, 96(45), 2017.
- L. Breiman. Random forests. Machine learning, 45:5–32, 2001.
- L. Breiman. Manual on setting up, using, and understanding random forests v3. 1. Statistics Department University of California Berkeley, CA, USA, 1(58):3–42, 2002.
- V. Brezina and G. Pallotti. Morphological complexity in written 12 texts. *Second language research*, 35(1):99–119, 2019.
- T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.
- M. Brysbaert and B. New. Moving beyond kučera and francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for american english. *Behavior research methods*, 41(4):977–990, 2009.
- B. J. Bushee, I. D. Gow, and D. J. Taylor. Linguistic complexity in firm disclosures: Obfuscation or information? *Journal of Accounting Research*, 56(1): 85–121, 2018.
- G. Çetinkaya. Türkçe metinlerin okunabilirlik düzeylerinin tanımlanması ve sınıflandırılması. 2010.
- T. Chen, T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, K. Chen, R. Mitchell, I. Cano, T. Zhou, et al. Xgboost: extreme gradient boosting. *R* package version 0.4-2, 1(4):1–4, 2015.
- X. Chen and D. Meurers. Word frequency and readability: Predicting the text-level readability with a lexical-level attribute. *Journal of Research in Reading*, 41(3):486–510, 2018.
- K. Collins-Thompson. Computational assessment of text readability: A survey of current and future research. *ITL-International Journal of Applied Linguistics*, 165(2):97–135, 2014.
- M. A. Covington and J. D. McFall. Cutting the gordian knot: The moving-average type-token ratio (mattr). *Journal of quantitative linguistics*, 17(2):94–100, 2010.

- E. Dale and J. S. Chall. A formula for predicting readability: Instructions. *Educational research bulletin*, pages 37–54, 1948.
- F. Dell'Orletta, S. Montemagni, and G. Venturi. Read-it: Assessing readability of italian texts with a view to text simplification. In *Proceedings of the second* workshop on speech and language processing for assistive technologies, pages 73–83, 2011.
- T. Deutsch, M. Jasbi, and S. Shieber. Linguistic features for readability assessment. *arXiv preprint arXiv:2006.00377*, 2020.
- J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.
- W. H. DuBay. Smart Language: Readers, Readability, and the Grading of Text. ERIC, 2007.
- L. Feng, N. Elhadad, and M. Huenerfauth. Cognitively motivated features for readability assessment. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 229–237, 2009.
- L. Feng, M. Jansche, M. Huenerfauth, and N. Elhadad. A comparison of features for automatic readability assessment. 2010.
- Y. Feng, J. Qiang, Y. Li, Y. Yuan, and Y. Zhu. Sentence simplification via large language models. arXiv preprint arXiv:2302.11957, 2023.
- R. Flesch. A new readability yardstick. Journal of applied psychology, 32(3):221, 1948.
- T. François and C. Fairon. An "ai readability" formula for french as a foreign language. In Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning, pages 466–477, 2012.
- T. François and E. Miltsakaki. Do nlp and machine learning improve traditional readability formulas? In *Proceedings of the First Workshop on Predicting and Improving Text Readability for target reader populations*, pages 49–57, 2012.
- E. Gibson. Linguistic complexity: Locality of syntactic dependencies. Cognition, 68(1):1–76, 1998.
- F. Gilardi, M. Alizadeh, and M. Kubli. Chatgpt outperforms crowd-workers for text-annotation tasks. arXiv preprint arXiv:2303.15056, 2023.

- I. Gonzalez-Dios, M. J. Aranzabe, A. D. de Ilarraza, and H. Salaberri. Simple or complex? assessing the readability of basque texts. In *Proceedings of COLING* 2014, the 25th international conference on computational linguistics: Technical papers, pages 334–344, 2014.
- R. Gunning. The fog index after twenty years. Journal of Business Communication, 6(2):3–13, 1969.
- A. Z. Guven. Readability of texts in textbooks in teaching turkish to foreigners. *The Anthropologist*, 18(2):513–522, 2014.
- J. Hale. Information-theoretical complexity metrics. Language and Linguistics Compass, 10(9):397–412, 2016.
- J. Hancke, S. Vajjala, and D. Meurers. Readability classification for german using lexical, syntactic, and morphological features. In *Proceedings of COLING 2012*, pages 1063–1080, 2012.
- M. Honnibal, I. Montani, S. Van Landeghem, A. Boyd, et al. spacy: Industrial-strength natural language processing in python. 2020.
- J. M. Imperial. Bert embeddings for automatic readability assessment. arXiv preprint arXiv:2106.07935, 2021.
- G. Jawahar, B. Sagot, and D. Seddah. What does bert learn about the structure of language? In ACL 2019-57th Annual Meeting of the Association for Computational Linguistics, 2019.
- K. Jeblick, B. Schachtner, J. Dexl, A. Mittermeier, A. T. Stüber, J. Topalis, T. Weber, P. Wesp, B. Sabel, J. Ricke, et al. Chatgpt makes medicine easy to swallow: An exploratory case study on simplified radiology reports. arXiv preprint arXiv:2212.14882, 2022.
- W. Jiao, W. Wang, J.-t. Huang, X. Wang, and Z. Tu. Is chatgpt a good translator? a preliminary study. arXiv preprint arXiv:2301.08745, 2023.
- S. Keklik. Türkçede 0-6 yaş çocuklarına öğretilmesi gereken en sık kullanılan 1200 kelime. Türkiye Sosyal Araştırmalar Dergisi, 3(3):1–28, 2010.
- J. P. Kincaid, R. P. Fishburne Jr, R. L. Rogers, and B. S. Chissom. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch, 1975.

- W. Kintsch and D. Vipond. Reading comprehension and readability in educational practice and psychological theory. *Perspectives on learning and memory*, pages 329–365, 2014.
- B. W. Lee and J. H.-J. Lee. Traditional readability formulas compared for english. arXiv preprint arXiv:2301.02975, 2023.
- B. W. Lee, Y. S. Jang, and J. H.-J. Lee. Pushing on text readability assessment: A transformer meets handcrafted linguistic features. arXiv preprint arXiv:2109.12258, 2021.
- J. Lee and S. Vajjala. A neural pairwise ranking model for readability assessment. arXiv preprint arXiv:2203.07450, 2022.
- P. Ley and T. Florio. The use of readability formulas in health care. *Psychology*, *Health & Medicine*, 1(1):7–28, 1996.
- H. H. Lim, T. Cai, J. S. Lee, and M. Liu. Robustness of hybrid models in cross-domain readability assessment. In *Proceedings of the The 20th Annual* Workshop of the Australasian Language Technology Association, pages 62–67, 2022.
- Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692, 2019.
- I. Loshchilov and F. Hutter. Fixing weight decay regularization in adam. 2017.
- X. Lu. Automatic analysis of syntactic complexity in second language writing. International journal of corpus linguistics, 15(4):474–496, 2010.
- B. Marşan, S. F. Akkurt, M. Şen, M. Gürbüz, O. Güngör, Ş. B. Özateş, S. Üsküdarlı, A. Özgür, T. Güngör, and B. Öztürk. Enhancements to the boun treebank reflecting the agglutinative nature of turkish. arXiv preprint arXiv:2207.11782, 2022.
- M. Martinc, S. Pollak, and M. Robnik-Šikonja. Supervised and unsupervised neural approaches to text readability. *Computational Linguistics*, 47(1):141–179, 2021.
- G. H. Mc Laughlin. Smog grading-a new readability formula. Journal of reading, 12(8):639–646, 1969.

- C. Meng, M. Chen, J. Mao, and J. Neville. Readnet: A hierarchical transformer framework for web article readability analysis. In Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part I 42, pages 33–49. Springer, 2020.
- A. Michail, S. Konstantinou, and S. Clematide. Uzh_clyp at semeval-2023 task 9: Head-first fine-tuning and chatgpt data generation for cross-lingual learning in tweet intimacy prediction. arXiv preprint arXiv:2303.01194, 2023.
- S. E. Petersen and M. Ostendorf. A machine learning approach to reading level assessment. *Computer speech & language*, 23(1):89–106, 2009.
- E. Pitler and A. Nenkova. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, pages 186–195, 2008.
- P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning. Stanza: A python natural language processing toolkit for many human languages. arXiv preprint arXiv:2003.07082, 2020.
- T. Smit, M. van Haastrecht, and M. Spruit. The effect of countermeasure readability on security intentions. *Journal of Cybersecurity and Privacy*, 1(4): 675–703, 2021.
- V. Sönmez. Metinlerin eğitselliğini saptamada matematiksel bir yaklaşım (sönmez modeli). Eğitim Araştırmaları, 10:24–39, 2003.
- K. Tanaka-Ishii, S. Tezuka, and H. Terada. Sorting texts by readability. *Computational linguistics*, 36(2):203–227, 2010.
- K. Tekşan, U. Süğümlü, and E. Çinpolat. Readability of turkish tales. Journal of Language and Linguistic Studies, 16(2):978–992, 2020.
- S. Tonelli, K. M. Tran, and E. Pianta. Making readability indices readable. In Proceedings of the First Workshop on Predicting and Improving Text Readability for target reader populations, pages 40–48, 2012.
- T. Turkben. Readability characteristics of texts in middle school turkish textbooks. Educational Policy Analysis and Strategic Research, 14(3):80–105, 2019.
- Turkish NLP Suite. $tr_core_n ews_t rf(revisionc7e06a7), 2022. URL.$

S. Vajjala. Trends, limitations and open challenges in automatic readability assessment research. *arXiv preprint arXiv:2105.00973*, 2021.

S. Vajjala and I. Lučić. Onestopenglish corpus: A new corpus for automatic readability assessment and text simplification. In *Proceedings of the thirteenth workshop* on innovative use of NLP for building educational applications, pages 297–304, 2018.

S. Vajjala and D. Meurers. On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the seventh workshop on building educational applications using NLP*, pages 163–173, 2012.

S. Vajjala and D. Meurers. Assessing the relative reading level of sentence pairs for text simplification. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 288–297, 2014.

L.-W. Wang, M. J. Miller, M. R. Schmitt, and F. K. Wen. Assessing readability formula differences with written health information materials: application, results, and recommendations. *Research in Social and Administrative Pharmacy*, 9(5):503–516, 2013.

S. Wang, Y. Liu, Y. Xu, C. Zhu, and M. Zeng. Want to reduce labeling cost? gpt-3 can help. *arXiv preprint arXiv:2108.13487*, 2021a.

Z. Wang, A. W. Yu, O. Firat, and Y. Cao. Towards zero-label language learning. *arXiv preprint arXiv:2109.09193*, 2021b.

Z. Weiss, X. Chen, and D. Meurers. Using broad linguistic complexity modeling for cross-lingual readability assessment. In *Proceedings of the 10th Workshop on NLP for Computer Assisted Language Learning*, pages 38–54, 2021.

T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45, 2020.

W. Xu, C. Callison-Burch, and C. Napoles. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297, 2015.