Exploring Hybrid Linguistic Features for Turkish Text Readability

Ahmet Yavuz Uluslu
Universität Zürich
ahmetyavuz.uluslu@uzh.ch

Gerold Schneider
Universität Zürich
gschneid@cl.uzh.ch

Abstract

This paper presents the first comprehensive study on automatic readability assessment of Turkish texts. We combine state-of-the-art neural network models with linguistic features at lexical, morphosyntactic, syntactic and discourse levels to develop an advanced readability tool. We evaluate the effectiveness of traditional readability formulas compared to modern automated methods and identify key linguistic features that determine the readability of Turkish texts.

1 Introduction

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, 2022). In the healthcare sector, 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). 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. Although manual linguistic analysis by domain experts provides valuable insights, it is time-consuming, costly and subject to individual interpretation, which can lead to variability and subjectivity in the annotation results (Deutsch et al.,

2020). Recent research in the field has focused on developing automated methods for extracting linguistic predictors and training models for readability assessment. Despite these crucial applications and developments, the readability efforts in Turkish have largely been confined to traditional readability formulas, such as Flesch-Kincaid (Kincaid et al., 1975) and its adaptations (Ateşman, 1997; Bezirci and Yilmaz, 2010; Çetinkaya, 2010). Several previous studies have pointed out the shortcomings of these formulas (Feng et al., 2010, 2009). They typically rely on superficial text features such as sentence length and word length. 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. In this paper, we present the first ARA study for Turkish. Our study combines traditional raw text features with lexical, morpho-syntactic, and syntactic information to create an advanced readability assessment tool for Turkish. We demonstrate the effectiveness of our tool on a new corpus of Turkish popular science magazine articles, published for different age groups and educational levels. Our study aims to contribute to the development of automated tools for accessibility, educational research, and language learning in Turkish.

The rest of the paper is organized as follows. In Section 2, we review related work on readability assessment and machine learning-based approaches. In Section 3, we describe our corpus and the linguistic features used in our study. In Section 4, we present the results of our experiments and analyze the effectiveness of our tool. Finally, in Section 5, we conclude our research and discuss future directions.

2 Previous Work

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. Initially developed for the English language, the Flesch Reading Ease formula required recalibration for its application to Turkish, a task undertaken by Ateşman (1997). 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 Çetinkaya (2010), which also assigned appropriate grade levels, thus facilitating its practical use in the Turkish educational context. Not long after the adaptation, Bezirci and Yilmaz (2010) introduced an important refinement, akin to the approach taken in the SMOG formula. They propose that features based on polysyllabic words and the total number of syllables present in the document provide distinct indications of text complexity. Accordingly, they included the counts of polysyllabic words (those with 3-, 4-, and 5+ syllables). Sönmez (2003) encountered inconsistencies when applying the Gunning FOG Index to Turkish texts which led to the development of their adaptation. The limitations are mainly due to the subjective nature of the formula in identifying complex words and concepts, which contrasts with other formulas that use easier-to-identify criteria such as syllable counts.

Readability assessment has found practical applications in several areas in Turkish, particularly in the fields of medicine and education. For instance, researchers have used the Flesch-Kincaid and Atesman readability formulae to assess the readability of anaesthesia consent forms in Turkish hospitals, which led to valuable insights into how these documents could be optimised 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 readability 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).

Readability formulae have inherent limitations that can affect their accuracy and applicability. Given the unique phonetic attributes, sentence formation patterns, and mean syllable length 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 levels of readability requirements (Akgül, 2019; Akgül, 2022) without accounting for the issues of calibration. Furthermore, the evolution of language over time may necessitate periodic re-calibration of these formulas (Lee and Lee, 2023). As language trends evolve and new words and phrases become more common, readability formulas must adapt to remain accurate and relevant. Previous research shows that traditional readability metrics perform unreliably 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 considerable text sample sizes for traditional measures introduces another impediment, even though the theoretical minimum size for a text sample has yet to be conclusively established. 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.

In recent years, research in ARA has shifted from traditional linear models, which use simple metrics such as word and sentence length to estimate the reading level of a text, to fine-grained features (Collins-Thompson, 2014). These features often include output of machine learning models trained on a combination of word counts, lexical patterns, discourse analysis, morphology, and syntactic structures. There has been an emerging trend toward using neural models for ARA. These models have demonstrated the capacity to implicitly capture the previously mentioned features without the need for manually-defined feature extraction (Jawahar et al., 2019). Martinc et al. (2021) and Imperial (2021) experimented with contextual embeddings of BERT (Devlin et al., 2019) for the readability assessment task, achieving par or better results than feature-based approaches. 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.

3 Corpus

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). 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 Magazines [1] spanning the period 2007 to 2022. The articles are openly published and 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 Meraklı Minik (for ages 0-6), Bilim Çocuk (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 (Vajjala and Meurers, 2012), 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.

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

---

[1] https://yayinlar.tubitak.gov.tr/

Level	Avg. Words	Std. Dev	Nr. of Articles
ELE	120.95	67.35	750
INT	154.99	93.57	750
ADV	327.08	187.54	750

Table 1: Descriptive Corpus Statistics

length may have a significant impact on determining the reading level of a given text.

We also performed a 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). 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.

Feature	ELE	INT	ADV
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

4 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[2] for the majority of general tasks, including entity recognition, POS tagging, and dependency parsing. We use the Stanford Stanza parser version 1.5.0 (Qi et al., 2020) for constituency parsing.

4.1 Traditional Features (TRAD)

Traditional or surface-based features are commonly used to predict the readability of Turkish texts, and

[2] https://huggingface.co/turkish-nlp-suite/tr_core_news_trf

we also adopt them as a baseline for our study. Specifically, we extract 7 traditional features, including Turkish adaptations of well-known readability formulas such as Atesman and Cetinkaya-Uzun, as well as average numbers of words and syllables per document. 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.

4.2 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 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.

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 the average and maximum depths of the constituency and dependency tree structures present in the text (Dell'Orletta et al., 2011).

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.

4.3 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).

Reading Level	Example
Elementary	Burası bir doğa koruma merkezi. Burada annesi ve babası olmayan turna yavruları var. Merkezde çalışanlardan biri özel bir giysi giyip koluna bir turna kuklası geçirmiş. *(This is a nature conservation centre. There are crane chicks without a mother and father. One of the workers at the centre wears a special suit and a crane puppet on his arm.)*
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. *(Robots are complex machines that do the jobs that humans do, instead of them. They are programmed to do these jobs. They fulfil certain commands automatically or by remote control.)*
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. *(The existence of magnetic perception in many living things is scientifically known. Bacteria, snails, frogs and lobsters can sense the Earth's magnetic field, migratory birds and sea turtles can navigate, and dogs can point out a hidden bar magnet when trained to do so.)*

Table 3: Example sentences for three reading levels

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 extract noun, verb, adjective, and adverb variations, which represent the proportion of the respective category's words to the total.

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). 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.

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 extracted the average word and sentence frequency for both early and late-acquired words. We calculate features based on the average log10 values similar to the SubtlexUS corpus (Brysbaert and New, 2009).

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 textbook[3] 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.

4.4 Morphological features (MORPH)

Morphological complexity plays a significant role in readability assessment, particularly in languages that are morphologically richer than English such as German (Hancke et al., 2012) and Basque (Gonzalez-Dios et al., 2014). In our study, we integrate the Morphological Complexity Index (MCI) from Brezina and Pallotti (2019). The MCI cap-

---

[3] https://www.turkishtextbook.com/most-common-words/

tures 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. MCI 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 such as Turkish and Hungarian.

4.5 Discourse features (DISCO)

The final group of features we examine are entity density features. 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 shown by Feng et al. (2009, 2010).

5 Experiments

We experiment with four different setups: trad-baseline (non-neural model with shallow features), modern-baseline (non-neural model with linguistic features), neural (pretrained transformer models), and hybrid (modern-baseline + neural). We use 10-fold cross-validation (10FCV) and evaluate our models using standard metrics such as accuracy, precision, recall, and macro F1-score. Specifically, we choose traditional learning algorithms such as Logistic Regression, Support Vector Machines, Random Forest and XGBoost as our baseline models. We perform a randomised search to explore a reasonable range of hyper-parameter values. We apply a grid search to identify the optimal combination of hyper-parameter values within this range.

5.1 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 (trad-baseline) is designed to be on par with traditional readability formulas and is reliant on shallow linguistic features such as sentence and word lengths. 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. We expand our feature set and include a more diverse set of linguistic feature groups (modern-baseline). We are 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.

5.2 Neural Models

We extend the established usage of transformer-based models in readability assessment (Deutsch et al., 2020; Martinc et al., 2021; Lee et al., 2021) and opt for the BERTurk model[4] for our analysis. 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, 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 three epochs.

5.3 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 human-defined linguistic features, which have been key components in traditional non-neural models (Deutsch et al., 2020). Our hybrid model takes a straightforward approach similar to that of Imperial (2021) and Lee et al. (2021). It combines the soft label predictions generated by the neural model with handcrafted features. These features are then used as input to a non-neural (Random Forest) model.

6 Results

We compare the performance of traditional and modern baselines to illustrate the process of arriving at the best-performing model. The process of feature and model selection for the baseline models

---

[4]`https://huggingface.co/dbmdz/`
`bert-base-turkish-uncased`

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 4: Performance comparison (modern-baseline) of readability models

was carried out based on the results obtained from different combinations.

6.1 Baseline: Feature and Model Evaluation

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

Table 5: Incremental contribution of each feature to the RandomF model

Through evaluation of four distinct models, namely Support Vector Machines (SVM), Random Forest (RandomF), Logistic Regression (LogR), and XGBoost, we assessed combinations of five different linguistic groups: traditional (TRAD), lexico-semantic (LXSM), syntactic (SYNX), morphological (MORPH), and discourse (DISCO) features. Table 6 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%. Importantly, all of the linguistic groups used provide orthogonal or distinct information. Table 5 demonstrates how each contributing linguistic group incrementally improves the accuracy of the Random Forest model. Their combined strength ultimately achieves the highest overall accuracy score.

Model	Acc	Prec	Rec	F1
trad-baseline	65.7	67.5	66.8	66.7
modern-baseline	85.3	85.3	85.1	85.1
neural	92.8	93.1	92.6	92.8
hybrid	**96.1**	**96.1**	**95.6**	**95.8**

Table 6: Performance comparison of readability approaches

The varying levels of performance between dif-

ferent approaches is demonstrated in Table 6. The hybrid model, which combines the strengths of both traditional and neural methodologies, outperforms 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.

7 Discussion

7.1 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) as shown in Figure 1 and 2.
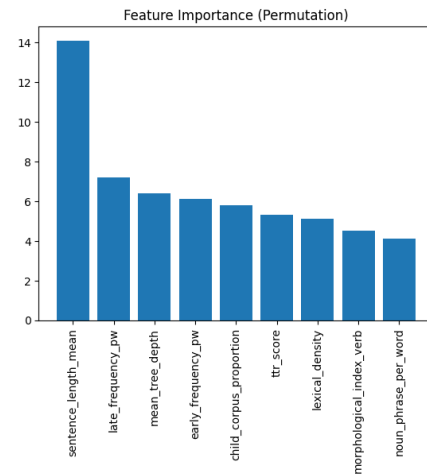

Figure 1: Feature importance by permutation on full model

7.2 Feature Correlation

We also considered model-independent analysis through Spearman correlation to gain additional perspective into the importance of features with respect to readability levels. Table 7 presents the ten features with the highest Spearman correlation
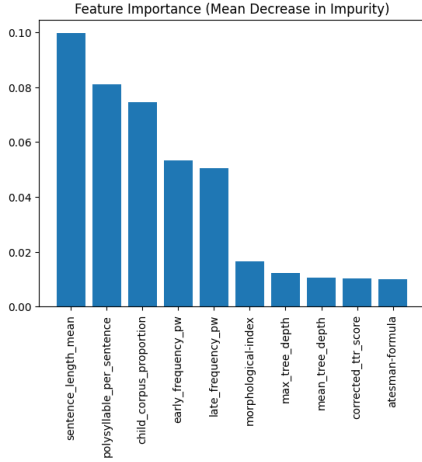
Figure 2: Feature Importance importance by MDI on full model

coefficients highlighting the significance of readability assessment.

Group	Feature	ρ
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 7: Top ten features ranked by their Spearman correlation coefficients

7.3 Lingustic Features

The analysis of feature importance consistently highlights the significant role of simple measures such as average sentence length and polysyllable counts. These findings align with previous research, where it has been shown that even compared to more complex feature extraction methods, a simple measure such as 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 evident from the performance improvement observed when including LXSM linguistic feature set in the modern-baseline method. It indicates that while traditional features are indeed valuable, incorporating fine-grained in-

formation at the semantic and lexical level can lead to an even better understanding of overall readability. 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.

8 Conclusion

We introduced a new readability corpus based on popular science magazine articles, providing a valuable resource for future research in Turkish readability assessment. 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 emphasise the importance of incorporating fine-grained linguistic features, as they provide more comprehensive insights into the complexity of Turkish texts. We showed the potential of hybrid models that combine fine-grained features with neural models by leveraging the strengths of both linguistic features and state-of-the-art transformers.

Acknowledgements

References

Elif Ahsen Acar, Deniz Zeyrek, Murathan Kurfalı, and Cem Bozşahin. 2016. 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, Portorož, Slovenia. European Language Resources Association (ELRA).

Yakup Akgül. 2019. 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.

Yakup Akgül. 2022. 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.

Ender Ateşman. 1997. Türkçede okunabilirliğin ölçülmesi. *Dil Dergisi*, 58(71-74).

Burak Bezirci and Asım Egemen Yilmaz. 2010. Metinlerin okunabilirliğinin ölçülmesi üzerine bir yazilim kütüphanesi ve türkçe için yeni bir okunabilirlik ölçütü. *Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik Dergisi*, 12(3):49–62.

Nilay Boztas, Dilek Omur, Sule Ozbılgın, Gözde Altuntas, Ersan Piskin, Sevda Ozkardesler, and Volkan Hanci. 2017. Readability of internet-sourced patient education material related to "labour analgesia". *Medicine*, 96(45).

Nilay Boztaş, Şule Özbilgin, Elvan Öçmen, Gözde Altuntaş, Sevda Özkardeşler, Volkan Hancı, and Ali Günerli. 2014. Evaluating the readability of informed consent forms available before anaesthesia: a comparative study. *Turkish journal of anaesthesiology and reanimation*, 42(3):140.

Vaclav Brezina and Gabriele Pallotti. 2019. Morphological complexity in written l2 texts. *Second language research*, 35(1):99–119.

Marc Brysbaert and Boris New. 2009. 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.

Brian J Bushee, Ian D Gow, and Daniel J Taylor. 2018. Linguistic complexity in firm disclosures: Obfuscation or information? *Journal of Accounting Research*, 56(1):85–121.

Gökhan Çetinkaya. 2010. Türkçe metinlerin okunabilirlik düzeylerinin tanımlanması ve sınıflandırılması. *Ankara University, PhD Thesis*.

Kevyn Collins-Thompson. 2014. Computational assessment of text readability: A survey of current and future research. *ITL-International Journal of Applied Linguistics*, 165(2):97–135.

Michael A Covington and Joe D McFall. 2010. Cutting the gordian knot: The moving-average type–token ratio (mattr). *Journal of quantitative linguistics*, 17(2):94–100.

Edgar Dale and Jeanne S Chall. 1948. A formula for predicting readability: Instructions. *Educational research bulletin*, pages 37–54.

Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2011. 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.

Tovly Deutsch, Masoud Jasbi, and Stuart Shieber. 2020. Linguistic features for readability assessment. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–17, Seattle, WA, USA → Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

William H DuBay. 2007. *Smart Language: Readers, Readability, and the Grading of Text.* ERIC.

Lijun Feng, Noémie Elhadad, and Matt Huenerfauth. 2009. Cognitively motivated features for readability assessment. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 229–237, Athens, Greece. Association for Computational Linguistics.

Lijun Feng, Martin Jansche, Matt Huenerfauth, and Noémie Elhadad. 2010. A comparison of features for automatic readability assessment. In *23rd International Conference on Computational Linguistics (COLING 2010), Poster Volume*, pages 276–284.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.

Thomas François and Cédrick Fairon. 2012. 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, Jeju Island, Korea. Association for Computational Linguistics.

Itziar Gonzalez-Dios, María Jesús Aranzabe, Arantza Díaz de Ilarraza, and Haritz Salaberri. 2014. 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, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Robert Gunning. 1969. The fog index after twenty years. *Journal of Business Communication*, 6(2):3–13.

Ahmet Zeki Guven. 2014. Readability of texts in textbooks in teaching turkish to foreigners. *The Anthropologist*, 18(2):513–522.

Julia Hancke, Sowmya Vajjala, and Detmar Meurers. 2012. Readability classification for German using lexical, syntactic, and morphological features. In *Proceedings of COLING 2012*, pages 1063–1080, Mumbai, India. The COLING 2012 Organizing Committee.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, Adriane Boyd, et al. 2020. spacy: Industrial-strength natural language processing in python.

Joseph Marvin Imperial. 2021. Bert embeddings for automatic readability assessment. *arXiv preprint arXiv:2106.07935*.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

S Keklik. 2010. Türkçede 0-6 yaş çocuklarına öğretilmesi gereken en sık kullanılan 1200 kelime. *Türkiye Sosyal Araştırmalar Dergisi*, 3(3):1–28.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. 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.

Walter Kintsch and Douglas Vipond. 2014. Reading comprehension and readability in educational practice and psychological theory. *Perspectives on learning and memory*, pages 329–365.

Bruce W. Lee, Yoo Sung Jang, and Jason Lee. 2021. Pushing on text readability assessment: A transformer meets handcrafted linguistic features. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10669–10686, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Bruce W Lee and Jason Hyung-Jong Lee. 2023. Traditional readability formulas compared for english. *arXiv preprint arXiv:2301.02975*.

Philip Ley and Tony Florio. 1996. The use of readability formulas in health care. *Psychology, Health & Medicine*, 1(1):7–28.

Xiaofei Lu. 2010. Automatic analysis of syntactic complexity in second language writing. *International journal of corpus linguistics*, 15(4):474–496.

Matej Martinc, Senja Pollak, and Marko Robnik-Šikonja. 2021. Supervised and unsupervised neural approaches to text readability. *Computational Linguistics*, 47(1):141–179.

G Harry Mc Laughlin. 1969. Smog grading-a new readability formula. *Journal of reading*, 12(8):639–646.

Sarah E Petersen and Mari Ostendorf. 2009. A machine learning approach to reading level assessment. *Computer speech & language*, 23(1):89–106.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*.

Tim Smit, Max van Haastrecht, and Marco Spruit. 2021. The effect of countermeasure readability on security intentions. *Journal of Cybersecurity and Privacy*, 1:675–704.

Veysel Sönmez. 2003. Metinlerin eğitselliğini saptamada matematiksel bir yaklaşım). *Marmara University, Master's Thesis*, 10:24–39.

Keziban Tekşan, Üzeyir Süğümlü, and Enes Çinpolat. 2020. Readability of turkish tales. *Journal of Language and Linguistic Studies*, 16(2):978–992.

Sara Tonelli, Ke Tran Manh, and Emanuele Pianta. 2012. Making readability indices readable. In *Proceedings of the First Workshop on Predicting and Improving Text Readability for target reader populations*, pages 40–48, Montréal, Canada. Association for Computational Linguistics.

Tuncay Turkben. 2019. Readability characteristics of texts in middle school turkish textbooks. *Educational Policy Analysis and Strategic Research*, 14(3):80–105.

Sowmya Vajjala. 2022. Trends, limitations and open challenges in automatic readability assessment research. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5366–5377, Marseille, France. European Language Resources Association.

Sowmya Vajjala and Ivana Lučić. 2018. 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, New Orleans, Louisiana. Association for Computational Linguistics.

Sowmya Vajjala and Detmar Meurers. 2012. 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, Montréal, Canada. Association for Computational Linguistics.

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

Zarah Weiss, Xiaobin Chen, and Detmar Meurers. 2021. 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, Online. LiU Electronic Press.

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