Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 5366–5377
Marseille, 20-25 June 2022

Trends, Limitations and Open Challenges in Automatic Readability Assessment Research

Sowmya Vajjala
National Research Council, Canada
sowmya.vajjala@nrc-cnrc.gc.ca

Abstract

Readability assessment is the task of evaluating the reading difficulty of a given piece of text. This article takes a closer look at contemporary NLP research on developing computational models for readability assessment, identifying the common approaches used for this task, their shortcomings, and some challenges for the future. Where possible, the survey also connects computational research with insights from related work in other disciplines such as education and psychology.

Keywords: readability assessment, survey, resources and evaluation

1. Introduction

Automatic Readability Assessment (ARA) refers to the task of modeling the reading and comprehension difficulty of a given piece of text, for a given target audience. This has a broad range of applications in both machine facing and human facing scenarios. Some examples of human facing scenarios are: choosing appropriate reading materials for language teaching (Collins-Thompson and Callan, 2004), supporting readers with learning disabilities (Rello et al., 2012) and self-directed learning (Beinborn et al., 2012). In machine facing scenarios, ARA is used in scenarios such as for ranking search results by their reading level (Kim et al., 2012), generating translated text controlled for reading level (Marchisio et al., 2019; Agrawal and Carpuat, 2019), and evaluating automatic text simplification (Alva-Manchego et al., 2020). TextEvaluator ™[1], used to determine whether a reading material is appropriate for a grade level in classroom instruction, is a well known real world application of ARA. Apart from such uses around and within the field of NLP, the general idea of readability assessment is used in a range of other scenarios. A common case is of medical research, where it was used for assessing patient education materials (Sare et al., 2020) and consent forms (Perni et al., 2019; Lyatoshinsky et al., 2019), for example. This broad application range highlights ARA as one of the important applications of NLP.

Research into measuring how difficult (or easy) is a text to read is now a century old (e.g., Thorndike (1921), Lively and Pressey (1923), Vogel and Washburne (1928)). Early research focused on creating lists of difficult words and/or developing a readability "formula", which is a simple weighted linear function of easy to calculate variables such as number/length of words/sentences in a text, percentage of difficult words etc. This resulted in several readability formulas such as Flesch Reading Ease (Flesch, 1948), SMOG (McLaughlin, 1969), Dale-Chall readability formula (Dale and Chall, 1948) etc. (see Dubay (2007) for a detailed survey of such formulae).

NLP researchers started taking interest in this problem only in the past two decades. From statistical language models and feature engineering based machine learning approaches to more recent deep neural networks, a range of approaches have been explored so far for this task. Despite this, a lot of application scenarios involving the use of ARA rely on traditional formulae even within NLP. For example, Marchisio et al. (2019) uses the "traditional formulae" such as Dale-Chall, Flesch Reading ease etc. as a measure of readability to control the reading level of machine translated text. In the scenarios outside of NLP, such as the use cases in medical research mentioned earlier too, one would notice the strong domination of traditional formulae. Possible reasons for this situation could be a lack of awareness of the state of the art in ARA or difficulty in using and interpreting it easily for their purpose.

Analyzing the reasons for this scenario would require taking a closer look at current methods in ARA research to understand the limitations in its adaptability. To our knowledge, there has only been one comprehensive ARA survey (Collins-Thompson, 2014) so far. There have been a lot of newer approaches to ARA since then, and researchers in other disciplines such as education have also published their perspectives on validation and evaluation of ARA approaches (e.g., Hiebert and Pearson (2014)). Further, the approach of the previous survey was also oriented more towards NLP researchers working on ARA. In this background, this paper aims to take a fresh look at ARA considering inputs from other disciplines where needed, and also cover recent research on various aspects of ARA, to get a generalized and contemporary picture about this NLP task.

The paper starts with an overview of the topic (Sections 1 and 2) and summarizes contemporary ARA research in NLP by identifying some common trends (Section 3). It then discusses their shortcomings (Section 4) in an attempt to understand why this large body of research is not reflected in its usage in various ap-

[1] `https://textevaluator.ets.org/TextEvaluator/`

plication scenarios. Finally, it identifies some challenges for future research (Section 5). Where possible, insights from other disciplines is summarized as well. Note that the terms readability and text complexity are used interchangeably in this paper, as is common in NLP research, although one can see more fine grained difference between the usage of these terms in education or psychology literature (e.g., Valencia et al. (2014)).

This survey is potentially useful for three kinds of readers:

1. NLP Researchers specifically working on ARA and other related problems (e.g., text simplification) may find this survey useful to understand the task holistically and identify language specific challenges.

2. Other NLP researchers can get a general overview of ARA and how to incorporate it into their systems.

3. Researchers from other disciplines looking to use ARA for their research can get an overview of the state of research in the field and what they can use easily.

2. Related Work

While there was been a lot of work in the NLP community on developing computational models for readability assessment across languages, there has not been much work synthesizing this research. Collins-Thompson (2014) is the most recent, comprehensive survey on this topic, to our knowledge. It gave a detailed overview of the various approaches to ARA and identified the development of user-centric models, data driven measures that can be easily specialized to new domains, and the inclusion of domain/conceptual knowledge into existing models as some of the potential research directions for future. François (2015) presented a historical overview of readability assessment focusing on early research on traditional formulae and identified three challenges for future work - validity of the training data, developing ARA approaches for different domains, and difficulty in estimating readability at different granularities (e.g., words and sentences).

Outside of NLP, Nelson et al. (2012) compared and evaluated a few existing proprietary text difficulty metrics (for English) using a range of reading difficulty annotated corpora and assessed the implications of such measures for education. In 2014, the Elementary School Journal published a special issue on understanding text complexity (Hiebert and Pearson, 2014), which offered multi-disciplinary perspectives on various aspects of ARA and its implications to education. Concluding that readability involves dimensions other than text and much more research is needed on the topic, the special issue cautioned about the danger of focusing on text readability scores alone. While the last two are not survey articles per se, they are included here as they summarize the findings from research that is not common knowledge in NLP research on ARA.

In the current survey, the focus is more on the recent developments in ARA research in NLP, drawing inputs from existing body of research in other related disciplines as needed. The goal of this paper is to provide a general overview of the trends in research and not to have an exhaustive listing of all published research on this topic during this period. The paper aims to remain language agnostic in this study, focusing primarily on the approaches taken for corpus creation, modeling and evaluation.

3. Current Trends in ARA Research in NLP

ARA is generally modeled as a supervised machine learning problem in NLP literature. Hence, a typical ARA approach follows the pipeline depicted in Figure 1.

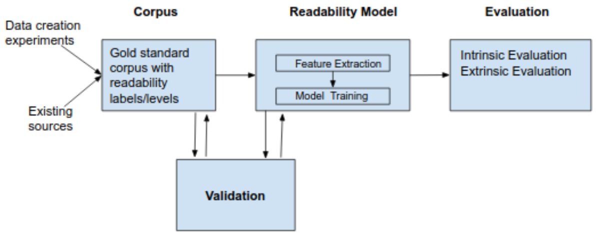

Figure 1: Typical ARA pipeline

ARA approaches rely on a gold standard training corpus annotated with labels indicating reading level categories, or numbers indicating a graded scale (**Corpus**). As with any machine learning problem, the next step consists of feature extraction and training a model (**Readability Model**). The final step in this process is an evaluation of the effectiveness of the model (**Evaluation**). A not so commonly seen, but essential step in this process is **Validation**, which evaluates not the model, but the process itself, including the corpora and features. Rest of this section discusses each of these steps in detail by giving an overview of representative approaches taken by researchers in handling these stages of ARA, and what changed in the recent few years, compared to Collins-Thompson (2014)'s survey.

3.1. Corpus

Training data in ARA comes from various sources. They can be broadly classified into two categories: expert annotated and non-expert annotated. Textbooks, or other graded readers carefully prepared by trained authors targeting audience at specific grade levels can be

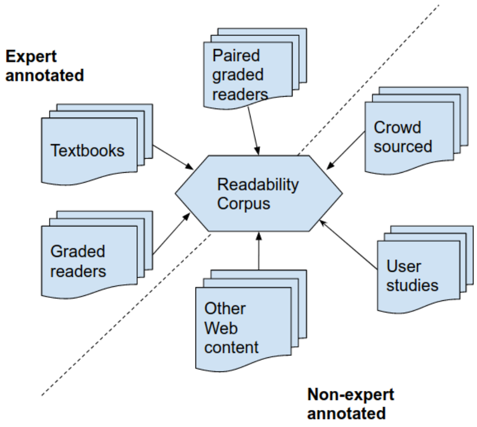

Figure 2: Various forms of ARA Corpora

termed as "expert annotated". These are the most common forms of training data seen in ARA research. On the other hand, some ARA work also relied on available web content, or doing crowd sourcing experiments and user studies to collect data. In such cases, we either do not know who did the annotations or we are getting them from a target reader population, who need not be experts on the linguistic aspects of readability. Figure 2 summarizes the different forms of data sources in ARA research, and the rest of this section discusses each of them in detail.

Textbooks: Textbooks have been a common source of training data for ARA research, when available, for several languages such as English (Heilman et al., 2007), Japanese (Sato et al., 2008), German (Berendes et al., 2018), Swedish (Pilán et al., 2016), French (François and Fairon, 2012), Bangla (Islam et al., 2012) and Greek (Chatzipanagiotidis et al., 2021), to name a few. They are considered to be naturally suited for ARA research as one would expect the linguistic characteristics of texts to become more complex as school grade increases. On a related note, Xia et al. (2016) collected reading comprehension passages from language exams conducted at different proficiency levels for building ARA models.

However, it is not always possible to have a readily accessible dataset of textbooks, as many textbooks are also under copyright and may not be accessible in a digitized form. Thus, most of the above mentioned corpora are not available for other researchers, which makes them a valuable, but not viable data source. A closer alternative is to use graded readers.

Graded Readers: This paper refers to non-textbook reading materials prepared by teachers or other experts, which are separated into some categorization of reading levels, as graded readers. Typically, these materials are derived from sources such as: news articles rewritten to suit the target reading level, encyclopedia articles written for adults and children separately, or children's readers from book publishing companies. WeeBit (Vajjala and Meurers, 2012) is one of the widely used graded reader corpus used for English ARA. Such graded readers exist for other languages as well. For example, Imperial (2021) recently described one such dataset for Filipino language.

In the recent past, corpora such as Newsela (Xu et al., 2015) and OneStopEnglish (Vajjala and Lučić, 2018) were created for English, which can be called **Paired Graded Readers**. Instead of having a collection of unrelated documents at each reading level, these corpora have the same documents rewritten to suit different reading levels. Newsela corpus, which also has a Spanish subset, was used to build text simplification systems (Štajner and Nisioi, 2018) and generating machine translated text at varying reading levels (Agrawal and Carpuat, 2019) in the past.

Other web content: When there are no available texts annotated with reading level, it is common to find other documents from the web which have some form of inherent reading level grouping. Simple Wikipedia[2] was widely used along with Wikipedia to build a easy versus difficult ARA system for English (Napoles and Dredze, 2010). A sentence aligned version of this dataset was also used for automatic text simplification (Hwang et al., 2015).

Other such websites have been used in other ARA approaches for English (Vajjala and Meurers, 2013), German (Hancke et al., 2012), Italian (Dell'Orletta et al., 2011) and Basque (Gonzalez-Dios et al., 2014) among others. (Azpiazu and Pera, 2019) used Vikidia[3] together with Wikipedia to compile a multilingual readability dataset in 7 languages.

Taking a slightly different approach, Eickhoff et al. (2011) relied on the topic hierarchy in Open Directory Project to group web articles based on whether they are appropriate to a certain age group. Vajjala and Meurers (2014a) used a corpus of TV program subtitles grouped into three age groups, collected from BBC channels. In the absence of readily available corpora annotated with reading levels, this seems to be the most common way of procuring some form of leveled text corpus for this task.

Crowdsourcing: All the above mentioned approaches relied on some form of an existing data source suitable for training ARA models. De Clercq et al. (2014) described the usefulness of a crowdsourcing for ARA, where non-expert readers/general public are shown two unrelated texts (in Dutch) each time and are asked to compare them in terms of their reading difficulty. Comparing these judgments with expert (e.g., teacher) judgments, they concluded that crowdsourcing is a viable alternative to expert annotations for this task.

User studies: Another way to gather an ARA corpus is by conducting user studies. For example, vor der

[2] https://simple.wikipedia.org/

[3] https://www.vikidia.org/

Brück et al. (2008) conducted a user study with 500 German documents from municipal domain, and non-expert readers were asked to rate the texts on a 7 point Likert scale (Likert, 1932) and used it to construct an ARA model. Similarly, Pitler and Nenkova (2008) conducted a user study where college students were asked to rate WSJ news articles on a scale, which was then used to build a readability model.

Štajner et al. (2017) collected user judgements of sentence level text complexity in the context of text simplification, for original, manually and automatically simplified sentences. Some studies conducted such studies to gather expert annotations as well. For example, Kate et al. (2010) described a dataset collected through a user study, rated separately by experts and naive readers. Shen et al. (2013) used a dataset collected and annotated by experts, in four languages - Arabic, Dari, English, and Pashto. Note that this is different from using available textbooks or graded readers, which are also graded by experts. User studies are typically conducted specifically for this task, and not for a generic use as in the case of other expert annotated resources.

Eye tracking and reading time information were also used in the past to annotate readability datasets (Nishikawa et al., 2013; Yaneva et al., 2015), which were done with less number of participants than the other user studies mentioned above. However, overall, user studies are not a common mode of corpus creation for ARA, owing to the time and effort involved. They also typically result in smaller datasets compared to other approaches for this task.

Among these different forms of resources, excepting paired graded readers and very few cases from "other web content", the texts/vocabulary at different reading levels in the corpora don't necessarily deal with the same content. For example, in the WeeBit corpus (Vajjala and Meurers, 2012), one of the commonly used corpus for English, articles tagged with different reading levels don't share the same topical content. As we will see in the next subsection, a majority of ARA models do not particularly control for topic variation. This leads us to question what the ARA models learn - is it a notion of text complexity, or topical differences among texts?

Further, whether these corpora are validated to be appropriate for the target audience is another important concern, not typically addressed in ARA research. For example, Simple Wikipedia is written for children and adults learning English. However, there is no evidence in the form of a user study that shows that this is indeed the case. Yet, it is used, along with Wikipedia, as a common data source for building readability models. Recently, Vajjala and Lucic (2019)'s study with over 100 participants concluded that the reading level annotations for texts in a paired graded corpus did not have any effect on reader's comprehension. In this background, an obvious question that arises is - what is the right corpus for this problem? François (2015) too discussed the issue of validity of training data in the context of ARA and called for more work in this direction. Another potential problem with existing ARA datasets is that of inter-annotator agreement. With user studies and crowd sourcing based data collection, it is possible to gather such information. However, we have no means of acquiring this information for other texts, especially the expert annotated corpora. It could be hard to understand and identify the shortcomings of ARA approaches without having a clear picture of human agreement on the task.

Finally, an often ignored issue in the discussion around ARA datasets is the domain of the texts. Textbooks and news articles seem to be the most commonly used genre, although we see focused datasets on ARA for legal/government documents, literary pieces etc. There is some past research that looked into genre effect on ARA models (Nelson et al., 2012; Dell'Orletta et al., 2014) and on how to develop an unbiased model across genres (Sheehan et al., 2013). However, this is an essential, but under-explored aspect in ARA research so far.

To conclude, while there are many ways of creating corpora for ARA research, we don't have many freely available corpora covering different languages, topics, and target domains, and we don't have strongly validated corpora suited for this task. Compared to Collins-Thompson (2014)'s survey, we can say that not much has happened in terms of corpora creation in ARA, and many questions remain.

3.2. Readability Model

The second step in ARA pipeline is to build the readability model, which includes both the feature extraction/text representation as well as training an ARA model. Research into building readability models in the past two decades has primarily relied on language models and feature engineering based machine learning approaches. Like with other NLP tasks, recent approaches relied on neural network and deep learning approaches for this task.

Features that are expected to influence the readability of a text come in various forms, from some simple, easy to calculate numbers such as number of words per sentence to more complex ones involving the estimation of a discourse structure in the document. While some of the advanced linguistic features such as coherence and cohesion are potentially hard to extract automatically, shallow variants e.g., noun overlap between adjacent sentences, implemented in Coh-Metrix (Graesser et al., 2004) are commonly used as proxies. Similarly, different kinds of text embeddings, which capture some form of syntactic and semantic properties of texts, also do not need advanced linguistic processing such as parsing, coreference resolution etc. Hence, instead of grouping features based on linguistic categories, as is commonly done, they are grouped based on the amount of language processing required in this

paper.
Figure 3 shows a summary of different kinds of features used in ARA research with examples at each step.

Feature Engineering: Features such as word length (in characters/syllables), sentence length, usage of different forms of word lists (Chen and Meurers, 2018), language models (e.g., Petersen and Ostendorf (2009)), models of word acquisition (Kidwell et al., 2011), measures of morphological variation and complexity (Hancke et al., 2012; Chatzipanagiotidis et al., 2021), syntactic complexity (Heilman et al., 2007; Vajjala and Meurers, 2012), psycholinguistic processes (Howcroft and Demberg, 2017) and other attributes have been extensively used for developing ARA models across languages. Some features relying on advanced processing such as coreference resolution and discourse relations (Pitler and Nenkova, 2008; Feng et al., 2010) have also been explored in the past, more for English, and to some extent for other languages such as French (Todirascu et al., 2016). (Collins-Thompson, 2014) presents a comprehensive summary of different kinds of features used in ARA.
Some recent research focused on learning task specific embeddings (e.g., Cha et al. (2017), Jiang et al. (2018)) for ARA. Although not common, there has also been some work on modeling conceptual difficulty (Jameel et al., 2012). An often ignored aspect of ARA is the reader. Kim et al. (2012) is one of rare works related to ARA which considers reader attributes such as interests, language level etc. into their model to rank search results by their reading level. Although not directly about ARA, Knowles et al. (2016) explored the relationship between a word comprehension and a learner's native language. More recently, Gooding et al. (2021a) proposed a method to predict text readability from the reader's scrolling behavior. Overall, though ARA approaches are meant to be for real users in most of the cases, we don't see much work on modeling user features in relation to ARA.
Feature engineering based ARA approaches typically employ feature selection methods to choose a subset of features that best work for the task from a larger set. Apart from generic methods such as information gain, feature correlation etc., genetic algorithm based optimization methods were also explored for this task (De Clercq and Hoste, 2016). Although some papers report on "best performing features" for a given dataset, we don't have a clear consensus on what groups of features perform better across languages and dataset. In a recent work, Weiss et al. (2021) showed experimented with English and German ARA using a broad linguistic feature set and presented a study of what features are consistently useful for both languages, and what are not, for this task. More research in this direction is needed to gain a better understanding of a core set of useful linguistic features for ARA across languages.

Training: In terms of training methods used, ARA is generally modeled as a supervised learning problem, especially classification. It is, however, not uncommon to see it being modeled as regression (Vajjala and Meurers, 2014b) and ranking (Tanaka-Ishii et al., 2010; Ma et al., 2012; Lee and Vajjala, 2022). Heilman et al. (2008) compared different approaches to learn an ARA model and showed that ordinal regression is better suited for the task. Xia et al. (2016) showed that pair wise ranking approach may generalize better compared to classification. Unlike such approaches, Jiang et al. (2019) proposed a graph propagation based approach to ARA, which can potentially consider the inter-relationships between documents while modeling readability. Finally, while almost all of ARA research has been modeling it as a supervised learning problem, Martinc et al. (2021) and Ehara (2021) proposed unsupervised approaches to measuring text readability in the recent past.

Like other NLP research, ARA in the past two years has been dominated by neural network based architectures. For example, Mohammadi and Khasteh (2019) proposed a multilingual readability assessment model using deep reinforcement learning and Meng et al. (2020) proposed ReadNet, a hierarchical self attention based transformer model for ARA. Contemporary research also explored different ways of combining linguistic features with transformer models (Deutsch et al., 2020; Lee et al., 2021).

In general, most readability approaches have been shown to work for one language, or individual models were developed for each language. However, Azpiazu and Pera (2019; 2020) study the development of multilingual and cross-lingual approaches to ARA using deep learning architectures. Weiss et al. (2021) studied whether a common core of linguistic features would be useful across languages, and performed zero-shot cross-lingual evaluation between English and German using a large collection of linguistic features.

To summarize, we may notice that the past two decades of ARA research closely followed other areas of NLP i.e, traditional feature engineering based methods heavily dominated most of the previous research, whereas recent research seems to see more deep learning based approaches. Compared to the previous survey from 2014, most new research on ARA seems to have focused particularly on this aspect. Yet, there doesn't seem to be a clear consensus on what works for ARA across languages. While Lee et al. (2021) concluded that a combination of transformer architecture and linguistic features give a better performance, Weiss et al. (2021) showed zero shot cross lingual transfer with linguistic features alone. More recently, Lee and Vajjala (2022) proposed a neural pairwise ranking model, that showed good zero shot cross-lingual transfer with only BERT embeddings as the starting point. So, while deep learning has clearly been useful for ARA, linguistic features still seem to show strong results for languages with existing NLP tools such as POS taggers and syntactic parsers.

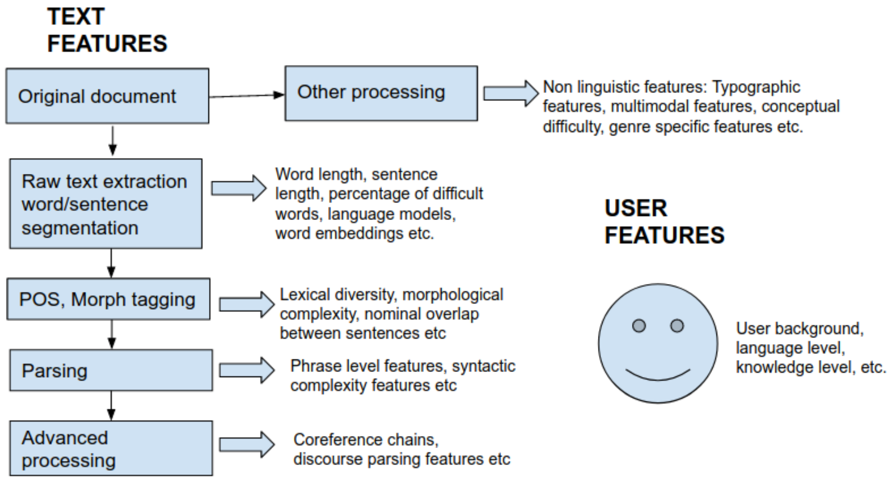
TEXT FEATURES
Original document
Other processing
Non linguistic features: Typographic features, multimodal features, conceptual difficulty, genre specific features etc.
Raw text extraction word/sentence segmentation
Word length, sentence length, percentage of difficult words, language models, word embeddings etc.
POS, Morph tagging
Lexical diversity, morphological complexity, nominal overlap between sentences etc
Parsing
Phrase level features, syntactic complexity features etc
Advanced processing
Coreference chains, discourse parsing features etc
USER FEATURES
User background, language level, knowledge level, etc.

Figure 3: Features used in ARA grouped by the amount of language processing needed

3.3. Evaluation

Evaluation of ARA can happen in two forms: intrinsic (evaluating on a standard test set) and extrinsic (evaluating on an end task). Most of the papers describing ARA models evaluate them intrinsically in terms of classification accuracy, F-score, Pearson/Spearman correlation (regression/ranking approaches), root mean square error (regression) and other such measures on held-out test data or in a cross-validated setup, as is conventionally done while evaluating supervised machine learning approaches. While it is not a default, we also see multi-corpus evaluation e.g., training on one corpus, testing on many; training and testing on many corpora (Nelson et al., 2012; Vajjala and Meurers, 2014b; Xia et al., 2016). Another way of evaluating if texts predicted by a ARA model to be "simple" result in better comprehension for the target reader group is through a user study. To our knowledge, such an evaluation has not been conducted so far for ARA models.

In terms of extrinsic evaluation, Pera and Ng (2012) and Kim et al. (2012) reported on experiments related to integrating readability approach into a search engine, and applying it for personalized search. Sheehan et al. (2014) deployed ARA models into a real-world tool. However, these examples are more of exceptions than norms, and such extrinsic evaluation is rare in ARA research, perhaps owing to the time and effort involved in such endeavours.

3.4. Validation

Validation is the step of assessing the accuracy of a process. Validation is distinct from evaluation as we are here evaluating other stages in model building, and not the ARA model itself. In the context of ARA research, validation is the step that answers the following two questions:

1. Are the reading level differences annotated in text corpora actually reflected in a reader's experience with the texts? i.e., Does the (annotated) reading level have any relation to reader comprehension?
2. Are the features used to represent a text theoretically valid, and can they reliably learn the reading level differences among texts?

Although these questions seem obvious, and have been posed many times in non-computational work on text readability in the past (e.g., Cunningham and Anne Mesmer (2014)), there is not much work in this direction in contemporary ARA research in NLP. Research related to TextEvaluator (Sheehan et al., 2014; Sheehan, 2017) has the only detailed analysis in this direction, to our knowledge. However, these are published outside of typical NLP venues, and hence, may not draw the attention of ARA researchers within NLP. François (2014) conducted a qualitative and quantitative analysis of a French as Foreign Language textbook corpus and concluded that there is a lack of consistent correlation among expert ratings, and that the texts assigned at the same level by the expert annotators showed significant differences in terms of lexical and syntactic features. Berendes et al. (2018) reached similar conclusions using a multidimensional corpus of graded German textbooks covering two school tracks and four publishers.

Although there are a few user studies aiming to study the relationship between readability annotations and reader comprehension (Crossley et al., 2014; Vajjala et al., 2016; Vajjala and Lucic, 2019), conclusions have been mixed. The most recent among these, Vajjala and Lucic (2019)'s study concluded that the reading level annotations for texts in a paired graded corpus did not have any effect on reader's comprehension.

To summarize, validation is an essential step in understanding whether our ARA models are really capturing the notion of text complexity, or just modeling randomly captured patterns in a given dataset. Clearly, there is not much work done on validation in ARA research, and this is an area which needs further work.

Now that we know about the trends in ARA research at different stages of building and evaluating a model, what is lacking?

4. Limitations

Based on this overview of current trends in the corpora creation, modeling, evaluation and validation of ARA, I identify the following limitations that are potentially preventing the adaption of modern ARA techniques into other research areas within and outside NLP.

1. **Multidimensional and Multimodal ARA models:** - Text readability involves several aspects of text, starting from typographical to linguistic, from conceptual difficulty to deeper pragmatics. However, contemporary ARA research tends to focus on the surface textual form. Topical or conceptual difficulty is not given much importance. Where it is considered, it is typically not combined with other aspects of readability.

   Further, texts don't exist in isolation. Many times, there is accompanying non-text data such as tables and/or images in the document. Although psycholinguists and cognitive psychologists explored such aspects through eye tracking studies in the past, I am are not aware of any research that touches upon these aspects in the context of NLP. To summarize, there is no framework yet (to our knowledge) that can incorporate a multidimensional, multimodal view of text complexity.
2. **Reader and Task considerations:** Research in education and psychology typically describes text complexity as a combination of text properties, reader (user) characteristics, and task complexity (Goldman and Lee, 2014; Valencia et al., 2014). However, within NLP, ARA research is almost always focused on text, with a small amount of research on reader modeling (Kim et al., 2012; Gooding et al., 2021a) and how what is complex can depend on a reader's language proficiency (Gooding et al., 2021b). While some research on modeling task complexity started to emerge (Kühberger et al., 2019), I am not aware of any approach that considers task complexity in the context of ARA or combine all the three aspects.
3. **Availability of corpus resources:** There is clearly a lot of work on ARA across many languages. Yet, we don't don't see a lot of publicly available corpora. Even when available, one has ask whether the corpora suit the target scenario. For example, one cannot use a corpus of textbooks to evaluate ARA models that intend to serve, say, dyslexic readers, as the reading difficulties experienced by dyslexic readers are completely different from first language readers learning subject matter in school. Similarly, it is not appropriate to use a corpus of news articles to develop a readability measure for legal texts. Such lack of available (and diverse) corpora can limit the development of ARA models tailored to specific application scenarios.
4. **Availability of ready to use tools:** There is not much of readily usable code artefacts related to building and using ARA models online. While some researchers shared code to reproduce their experiments (e.g., Ambati et al. (2016), Howcroft and Demberg (2017)), there is not much usable code for other NLP researchers or off the shelf tools for researchers from other disciplines. Recent tools such as LingFeat (Lee et al., 2021) provide implementations to a wide range of linguistic features, including traditional readability formulae, but don't have any ready to use pre-trained readability systems. Availability of such tools can potentially be useful for researchers from other disciplines wanting to use readability assessment approaches to answer research questions in their own domains.
5. **Lack of extrinsic evaluation:** Typically, ARA approaches are evaluated intrinsically, using cross validation or held out test set. It is rare to see an extrinsic evaluation when we consider a typical ARA research paper. This makes it particularly hard for practitioners to understand whether an approach works in an applied scenario.
6. **Lack of validation and interpretation:** The most common approach taken in building an ARA model is to take an available corpus, extract various kinds of features, and train different models and compare them. However, there is very little research on whether the corpus is suited for the task, whether the features themselves are actually useful, or if they have a theoretical grounding. Further, it is hard to understand what exactly does a model learn about text complexity. These issues make it difficult for researchers from other domains wanting to adapt modern ARA methods, and they instead turn to traditional formulae, which are relatively straight forward to interpret, even if they themselves are not validated either.

Although some of these limitations can be termed generic to NLP itself and not specific to ARA, this section attempted to highlight these issues in the context of contemporary ARA approaches. Among these, the first three limitations are of particular concern to NLP researchers, both in terms of using ARA in other NLP problems as well as furthering research on ARA itself. The remaining limitations are more general in nature, and would interest all the three target audience. I believe these are the factors that come between ARA research and its broader usefulness.

5. Challenges and Open Questions

In view of the above mentioned limitations and their potential consequences, I identify four major challenge areas where more future work is needed to address the current limitations of ARA.

1. **A framework to develop a holistic model of text readability**: We have seen that ARA research is primarily focused on textual features, especially those that focus on form. However, there are many other aspects such as conceptual difficulty, typographic features, user characteristics, task features etc, as we saw earlier. An obvious challenge would be to develop a unified model of ARA that encompasses all these aspects. However, it is not the work of one person or group, nor can it all be done in one go. So, an important first step in this direction (which can address limitations 1–2) would be to design an easily extendable framework to build a holistic model of readability by incrementally adding multiple dimensions, covering multi modal data. This would also necessitate the development of appropriate corpora and other resources suitable for this purpose.
2. **Models adaptable to new domain**: Any ARA model could still only be relevant to the target domain/audience and may not directly transfer to a new application scenario. Hence, approaches that can transfer an existing model into a new domain/audience should be developed. One potential avenue to explore in this direction is to model ARA as a ranking problem instead of classification or regression, as recent research concludes that it generalizes better than other models (Lee and Vajjala, 2022). This can address the limitation 3 mentioned earlier.
3. **Creation of open and diverse datasets and tools:** Development of openly accessible corpora which suit various application scenarios, for several languages is a major challenge in ARA research, as we saw earlier. New methods to quickly create (and validate) corpora need to be developed. Whether recent developments in data augmentation can be useful for developing ARA corpora is also something that can be explored in future. For widespread adaptation of research on ARA, and to progress towards a holistic model, ready to use tools should be developed. Tools such as Coh-Metrix (Graesser et al., 2011) and LingFeat[4] (Lee et al., 2021) that provide a range of linguistic features typically associated with readability assessment are a step in this direction. Along with these, tools that can show the predictions of ARA models should also be developed, to address the limitations 3–4.
4. **Developing Best Practices:** To support the creation of reusable resources (corpora/code) and to be able to reproduce/replicate results and understand SOTA, a set of best practices must be developed for ARA. Some inspiration for this can be drawn from the procedures and findings of the recently conducted REPROLANG challenge (Branco et al., 2020) which conducted a shared task to replicate some published NLP research. The best practices for ARA should also include guidelines for validating the corpora and features developed, as well as recommended procedures for developing interpretable approaches. This can help one address the limitations 5–6 to some extent. This will also potentially encourage non-NLP researchers to seriously consider employing more recent ARA models in their research. Some aspects of this challenge area (e.g., validation, interpretation) demand expertise beyond NLP methods and may require inter-disciplinary collaborations.

It has to be noted that some of these challenges are not necessarily specific to ARA, and are applicable across NLP in general. However, as with the previous section, this paper aims to discuss them in the context of ARA in particular and not in the context of entire NLP research. Further, This collection of ideas on challenges for future is by no means exhaustive, and I hope this survey initiates more discussion in this direction.

6. Conclusion

In this paper, I presented an overview of two decades of research on automatic readability assessment in NLP and connected it with related areas of research and applications. During this process I identified the limitations of contemporary research and identified some challenge areas for future. This analysis leads us to conclude that despite a large body of research, we don't yet have a clear picture of what are a good set of resources, modeling techniques that can be considered as SOTA across langauges in ARA. There is also a dearth of off the shelf tools and resources that support researchers and practitioners interested in ARA. Further, many challenges mentioned in previous surveys still remain. Considering that readability assessment has a wide range of applications in and outside NLP as it was seen from examples in Section 1, I think it is important to address these issues and enable the a broader adaption of ARA approaches within and outside NLP, over traditional formulae which only consider superficial aspects of language. More focus on validating NLP approaches to ARA, and on being able to interpret and relate model predictions to actual textual complexity may be the first steps in this direction.

Acknowledgements

I thank the anonymous reviewers for their detailed comments on the submitted version, Michel Simard and Rebecca Knowles for their comments on an earlier version of the paper, and Joseph Mervin Imperial, for actionable feedback before the final submission.

[4] `https://github.com/brucewlee/lingfeat`

7. Bibliographical References

Agrawal, S. and Carpuat, M. (2019). Controlling text complexity in neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1549–1564.

Alva-Manchego, F., Scarton, C., and Specia, L. (2020). Data-driven sentence simplification: Survey and benchmark. *Computational Linguistics*, 46(1):135–187.

Ambati, B. R., Reddy, S., and Steedman, M. (2016). Assessing relative sentence complexity using an incremental ccg parser. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1051–1057.

Azpiazu, I. M. and Pera, M. S. (2019). Multiattentive recurrent neural network architecture for multilingual readability assessment. *Transactions of the Association for Computational Linguistics*, 7:421–436.

Azpiazu, I. M. and Pera, M. S. (2020). Is cross-lingual readability assessment possible? *Journal of the Association for Information Science and Technology*, 71(6):644–656.

Beinborn, L., Zesch, T., and Gurevych, I. (2012). Towards fine-grained readability measures for self-directed language learning. In *Proceedings of the SLTC 2012 workshop on NLP for CALL; Lund; 25th October; 2012*, pages 11–19. Linköping University Electronic Press.

Berendes, K., Vajjala, S., Meurers, D., Bryant, D., Wagner, W., Chinkina, M., and Trautwein, U. (2018). Reading demands in secondary school: Does the linguistic complexity of textbooks increase with grade level and the academic orientation of the school track? *Journal of Educational Psychology*, 110(4):518.

Branco, A., Calzolari, N., Vossen, P., van Noord, G., Van Uytvanck, D., Silva, J., Gomes, L., Moreira, A., and Elbers, W. (2020). A shared task of a new, collaborative type to foster reproducibility: A first exercise in the area of language science and technology with reprolang2020. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5539–5545.

Cha, M., Gwon, Y., and Kung, H. (2017). Language modeling by clustering with word embeddings for text readability assessment. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 2003–2006.

Chatzipanagiotidis, S., Giagkou, M., and Meurers, D. (2021). Broad linguistic complexity analysis for greek readability classification. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 48–58.

Chen, X. and Meurers, D. (2018). Word frequency and readability: Predicting the text-level readability with a lexical-level attribute. *Journal of Research in Reading*, 41(3):486–510.

Collins-Thompson, K. and Callan, J. (2004). Information retrieval for language tutoring: An overview of the reap project. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 544–545.

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

Crossley, S. A., Yang, H. S., and McNamara, D. S. (2014). What's so simple about simplified texts? a computational and psycholinguistic investigation of text comprehension and text processing. *Reading in a Foreign Language*, 26(1):92–113.

Cunningham, J. W. and Anne Mesmer, H. (2014). Quantitative measurement of text difficulty: What's the use? *The Elementary School Journal*, 115(2):255–269.

Dale, E. and Chall, J. S. (1948). A formula for predicting readability: Instructions. *Educational research bulletin*, pages 37–54.

De Clercq, O. and Hoste, V. (2016). All mixed up? finding the optimal feature set for general readability prediction and its application to english and dutch. *Computational Linguistics*, 42(3):457–490.

De Clercq, O., Hoste, V., Desmet, B., Van Oosten, P., De Cock, M., and Macken, L. (2014). Using the crowd for readability prediction. *Natural Language Engineering*, 20(3):293–325.

Dell'Orletta, F., Montemagni, S., and Venturi, G. (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. Association for Computational Linguistics.

Dell'Orletta, F., Montemagni, S., and Venturi, G. (2014). Assessing document and sentence readability in less resourced languages and across textual genres. *ITL-International Journal of Applied Linguistics*, 165(2):163–193.

Deutsch, T., Jasbi, M., and Shieber, S. (2020). Linguistic features for readability assessment. *arXiv preprint arXiv:2006.00377*.

DuBay, W. H. (2007). *Unlocking language: The classic readability studies*. Impact Information.

Ehara, Y. (2021). Evaluation of unsupervised automatic readability assessors using rank correlations. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 62–72.

Eickhoff, C., Serdyukov, P., and De Vries, A. P. (2011). A combined topical/non-topical approach to identifying web sites for children. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 505–514.

Feng, L., Jansche, M., Huenerfauth, M., and Elhadad,

N. (2010). A comparison of features for automatic readability assessment. In *Coling 2010: Posters*, pages 276–284.

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

François, T. and Fairon, C. (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, July. Association for Computational Linguistics.

François, T. (2014). An analysis of a french as a foreign language corpus for readability assessment. In *Proceedings of the third workshop on NLP for computer-assisted language learning*, pages 13–32.

François, T. (2015). When readability meets computational linguistics: a new paradigm in readability. *Revue française de linguistique appliquée*, 20(2):79–97.

Goldman, S. R. and Lee, C. D. (2014). Text complexity: State of the art and the conundrums it raises. *the elementary school journal*, 115(2):290–300.

Gonzalez-Dios, I., Aranzabe, M. J., de Ilarraza, A. D., and Salaberri, H. (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.

Gooding, S., Berzak, Y., Mak, T., and Sharifi, M. (2021a). Predicting text readability from scrolling interactions. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 380–390, Online, November. Association for Computational Linguistics.

Gooding, S., Kochmar, E., Yimam, S. M., and Biemann, C. (2021b). Word complexity is in the eye of the beholder. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4439–4449, Online, June. Association for Computational Linguistics.

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

Graesser, A. C., McNamara, D. S., and Kulikowich, J. M. (2011). Coh-metrix: Providing multilevel analyses of text characteristics. *Educational researcher*, 40(5):223–234.

Hancke, J., Vajjala, S., and Meurers, D. (2012). Readability classification for german using lexical, syntactic, and morphological features. In *Proceedings of COLING 2012*, pages 1063–1080.

Heilman, M., Collins-Thompson, K., Callan, J., and Eskenazi, M. (2007). Combining lexical and grammatical features to improve readability measures for first and second language texts. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 460–467.

Heilman, M., Collins-Thompson, K., and Eskenazi, M. (2008). An analysis of statistical models and features for reading difficulty prediction. In *Proceedings of the third workshop on innovative use of NLP for building educational applications*, pages 71–79.

Hiebert, E. H. and Pearson, P. D. (2014). Understanding text complexity: Introduction to the special issue. *the elementary school journal*, 115(2):153–160.

Howcroft, D. M. and Demberg, V. (2017). Psycholinguistic models of sentence processing improve sentence readability ranking. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 958–968.

Hwang, W., Hajishirzi, H., Ostendorf, M., and Wu, W. (2015). Aligning sentences from standard wikipedia to simple wikipedia. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 211–217.

Imperial, J. M. (2021). Bert embeddings for automatic readability assessment. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 611–618.

Islam, Z., Mehler, A., and Rahman, R. (2012). Text readability classification of textbooks of a low-resource language. In *Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation*, pages 545–553.

Jameel, S., Lam, W., and Qian, X. (2012). Ranking text documents based on conceptual difficulty using term embedding and sequential discourse cohesion. In *2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, volume 1, pages 145–152. IEEE.

Jiang, Z., Gu, Q., Yin, Y., and Chen, D. (2018). Enriching word embeddings with domain knowledge for readability assessment. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 366–378.

Jiang, Z., Gu, Q., Yin, Y., Wang, J., and Chen, D. (2019). Graw+: A two-view graph propagation method with word coupling for readability assessment. *Journal of the Association for Information Science and Technology*, 70(5):433–447.

Kate, R. J., Luo, X., Patwardhan, S., Franz, M., Florian, R., Mooney, R. J., Roukos, S., and Welty, C. (2010). Learning to predict readability using diverse linguistic features. In *Proceedings of the 23rd international conference on computational linguistics*, pages 546–554. Association for Computational Linguistics.

Kidwell, P., Lebanon, G., and Collins-Thompson, K. (2011). Statistical estimation of word acquisition with application to readability prediction. *Journal of the American Statistical Association*, 106(493):21–30.

Kim, J. Y., Collins-Thompson, K., Bennett, P. N., and Dumais, S. T. (2012). Characterizing web content, user interests, and search behavior by reading level and topic. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 213–222.

Knowles, R., Renduchintala, A., Koehn, P., and Eisner, J. (2016). Analyzing learner understanding of novel l2 vocabulary. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 126–135.

Kühberger, C., Bramann, C., Weiß, Z., and Meurers, D. (2019). Task complexity in history textbooks: A multidisciplinary case study on triangulation in history education research. *History Education Research Journal*, 16(1):139–157.

Lee, J. and Vajjala, S. (2022). A neural pairwise ranking model for readability assessment. *Findings of the Association for Computational Linguistics: ACL 2022*, May.

Lee, B. W., Jang, Y. S., and Lee, J. H.-J. (2021). Pushing on text readability assessment: A transformer meets handcrafted linguistic features. *arXiv preprint arXiv:2109.12258*.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of psychology*.

Lively, B. A. and Pressey, S. L. (1923). A method for measuring the vocabulary burden of textbooks. *Educational administration and supervision*, 9(389-398):73.

Lyatoshinsky, P., Pratsinis, M., Abt, D., Schmid, H.-P., Zumstein, V., and Betschart, P. (2019). Readability assessment of commonly used german urological questionnaires. *Current urology*, 13(2):87–93.

Ma, Y., Fosler-Lussier, E., and Lofthus, R. (2012). Ranking-based readability assessment for early primary children's literature. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 548–552.

Marchisio, K., Guo, J., Lai, C.-I., and Koehn, P. (2019). Controlling the reading level of machine translation output. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, pages 193–203.

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

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

Meng, C., Chen, M., Mao, J., and Neville, J. (2020). Readnet: A hierarchical transformer framework for web article readability analysis. In *European Conference on Information Retrieval*, pages 33–49. Springer.

Mohammadi, H. and Khasteh, S. H. (2019). Text as environment: A deep reinforcement learning text readability assessment model. *arXiv preprint arXiv:1912.05957*.

Napoles, C. and Dredze, M. (2010). Learning simple wikipedia: A cogitation in ascertaining abecedarian language. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics and Writing: Writing Processes and Authoring Aids*, pages 42–50. Association for Computational Linguistics.

Nelson, J., Perfetti, C., Liben, D., and Liben, M. (2012). Measures of text difficulty: Testing their predictive value for grade levels and student performance. *Council of Chief State School Officers, Washington, DC*.

Nishikawa, H., Makino, T., and Matsuo, Y. (2013). A pilot study of readability prediction with reading time. In *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 78–84, Sofia, Bulgaria, August. Association for Computational Linguistics.

Pera, M. S. and Ng, Y.-K. (2012). Brek12: a book recommender for k-12 users. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 1037–1038.

Perni, S., Rooney, M. K., Horowitz, D. P., Golden, D. W., McCall, A. R., Einstein, A. J., and Jagsi, R. (2019). Assessment of use, specificity, and readability of written clinical informed consent forms for patients with cancer undergoing radiotherapy. *JAMA oncology*, 5(8):e190260–e190260.

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

Pilán, I., Vajjala, S., and Volodina, E. (2016). A readable read: Automatic assessment of language learning materials based on linguistic complexity. *arXiv preprint arXiv:1603.08868*.

Pitler, E. and Nenkova, A. (2008). Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, pages 186–195.

Rello, L., Saggion, H., Baeza-Yates, R., and Graells, E. (2012). Graphical schemes may improve readability but not understandability for people with dyslexia. In *Proceedings of the First Workshop on Predicting and Improving Text Readability for target reader populations*, pages 25–32. Association for Computational Linguistics.

Sare, A., Patel, A., Kothari, P., Kumar, A., Patel, N., and Shukla, P. A. (2020). Readability assessment of internet-based patient education materials related to treatment options for benign prostatic hyperplasia. *Academic Radiology*.

Sato, S., Matsuyoshi, S., and Kondoh, Y. (2008). Automatic assessment of japanese text readability based on a textbook corpus. In *LREC*.

Sheehan, K. M., Flor, M., and Napolitano, D. (2013). A two-stage approach for generating unbiased estimates of text complexity. In *Proceedings of the Workshop on Natural Language Processing for Improving Textual Accessibility*, pages 49–58.

Sheehan, K. M., Kostin, I., Napolitano, D., and Flor, M. (2014). The textevaluator tool: Helping teachers and test developers select texts for use in instruction and assessment. *The Elementary School Journal*, 115(2):184–209.

Sheehan, K. M. (2017). Validating automated measures of text complexity. *Educational Measurement: Issues and Practice*, 36(4):35–43.

Shen, W., Williams, J., Marius, T., and Salesky, E. (2013). A language-independent approach to automatic text difficulty assessment for second-language learners. In *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 30–38.

Štajner, S. and Nisioi, S. (2018). A detailed evaluation of neural sequence-to-sequence models for in-domain and cross-domain text simplification. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

Štajner, S., Ponzetto, S. P., and Stuckenschmidt, H. (2017). Automatic assessment of absolute sentence complexity. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence, IJCAI*, volume 17, pages 4096–4102.

Tanaka-Ishii, K., Tezuka, S., and Terada, H. (2010). Sorting texts by readability. *Computational linguistics*, 36(2):203–227.

Thorndike, E. L. (1921). *The teacher's word book*. Teacher's College, Columbia University.

Todirascu, A., François, T., Bernhard, D., Gala, N., and Ligozat, A.-L. (2016). Are cohesive features relevant for text readability evaluation? In *26th International Conference on Computational Linguistics (COLING 2016)*, pages 987–997.

Vajjala, S. and Lučić, I. (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.

Vajjala, S. and Lucic, I. (2019). On understanding the relation between expert annotations of text readability and target reader comprehension. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 349–359.

Vajjala, S. and Meurers, D. (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. Association for Computational Linguistics.

Vajjala, S. and Meurers, D. (2013). On the applicability of readability models to web texts. In *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 59–68.

Vajjala, S. and Meurers, D. (2014a). Exploring measures of "readability" for spoken language: Analyzing linguistic features of subtitles to identify age-specific tv programs. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, pages 21–29.

Vajjala, S. and Meurers, D. (2014b). Readability assessment for text simplification: From analysing documents to identifying sentential simplifications. *ITL-International Journal of Applied Linguistics*, 165(2):194–222.

Vajjala, S., Meurers, D., Eitel, A., and Scheiter, K. (2016). Towards grounding computational linguistic approaches to readability: Modeling reader-text interaction for easy and difficult texts. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, pages 38–48.

Valencia, S. W., Wixson, K. K., and Pearson, P. D. (2014). Putting text complexity in context: Refocusing on comprehension of complex text. *The Elementary School Journal*, 115(2):270–289.

Vogel, M. and Washburne, C. (1928). An objective method of determining grade placement of children's reading material. *The Elementary School Journal*, 28(5):373–381.

vor der Brück, T., Hartrumpf, S., and Helbig, H. (2008). A readability checker with supervised learning using deep indicators. *Informatica*, 32(4).

Weiss, Z., Chen, X., and Meurers, D. (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.

Xia, M., Kochmar, E., and Briscoe, T. (2016). Text readability assessment for second language learners. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 12–22.

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

Yaneva, V., Temnikova, I., and Mitkov, R. (2015). Accessible texts for autism: An eye-tracking study. In *Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility*, pages 49–57.