# Computational Assessment of Text Readability: A Survey of Current and Future Research

Running title: Computational Assessment of Text Readability

Kevyn Collins-Thompson
Associate Professor
University of Michigan, School of Information
105 South State St.
Ann Arbor, Michigan, U.S.A. 48109
Email: kevynct@umich.edu
Phone: +1 734-615-2132

Working draft

Last updated: Sept 8, 2014 10:36am

The author welcomes corrections, omissions, or comments sent to the above email address.

Abstract:

Assessing text readability is a time-honored problem that has even more relevance in today's information-rich world. This article provides background on how readability of texts is assessed automatically, reviews the current state-of-the-art algorithms in automatic modeling and predicting the reading difficulty of texts, and proposes new challenges and opportunities for future exploration not well-covered by current computational research.

Keywords: readability, reading difficulty, text complexity, computational linguistics, machine learning.

# Computational Assessment of Text Readability: A Survey of Current and Future Research

## 1. Introduction

For as long as people have originated, shared, and studied ideas through written language, the notion of text difficulty has been an important aspect of communication and education. As described by Zakaluk and Samuels (1988), scholars in ancient Athens more than two millennia ago noted a concern for text comprehensibility as part of the rhetorical training for law students: a legal argument or analysis was of little persuasive value if its audience could not understand it. Only within the last century, however, has a more systematic, scientific approach been taken to understanding the subjective and objective factors associated with text difficulty, and how best to support readers in their quest to understand more difficult texts, or find texts at the right level of difficulty.

As part of this systematic approach, *text readability* has been more formally defined as the sum of all elements in textual material that affect a reader's understanding, reading speed, and level of interest in the material (Dale & Chall, 1949). These elements may include features such as the complexity of sentence syntax; the semantic familiarity to the reader of the concepts being discussed; whether there is a supporting graphic or illustration; the sophistication of logical arguments or inference used to connect ideas; and many other important dimensions of content. In addition to text characteristics, a text's readability is also a function of the readers themselves: their educational and social background, interests and expertise, and motivation to learn, as well as other factors, can play a critical role in how readable a text is for an individual or population.

Given the importance of text readability in meeting people's information needs, along with modern access to ever-larger volumes of information, the implications of achieving effective text readability assessment are as diverse as the uses for text itself. The ability to quantify the readability of a text is achieved through the use of *readability measures* that take a text as input and estimate a numerical score or other form of prediction that indicates the level or degree of readability for a given population. In this survey, we focus less on the graphical aspects of readability, such as font size or color contrast, that affect a reader's initial ability to visually decode a text, and more on the linguistic features of a text that affect subsequent comprehension difficulty. Thus, we sometimes use the phrases *text difficulty* or *reading difficulty* synonymously with *text readability* for the purposes of this article.

Modern research on estimation of text readability, and the development of readability measures, has a history going back at least a century (cf. Chall, 1958). Yet far from being a 'solved' problem, automated assessment of text readability remains a challenging and highly relevant research area. Also notable is the key role that automated readability assessment can play in specific application domains where the accessibility of critical information is especially important and may currently be lacking. These include finding educational material of the right difficulty for students in textbooks and online; calibrating public and private health information so that it is understandable by

the general public and individual patients, in the form of medical instructions, questionnaires, pamphlets, online resources, and the like; producing effective product guides and other documentation; creating informative and easy-to-understand Web sites and forms for critical government services; and supporting the world's information needs via the Web and social media using search engines and recommender systems.

With the advent of increasingly sophisticated computation methods, along with new sources of data and applications to the Web and social media, the field of automated text readability assessment has evolved significantly in the last decade, and its utility and scope across applications have increased dramatically. On the one hand, widely-used traditional readability measures like Flesch-Kincaid, which estimate text readability based on simple functions of two or three linguistic variables such as syllable and word counts, have been used for decades on traditional texts. However, there is now a shift underway away from these simple but shallow traditional measures, in favor of data-driven, user-centric, knowledge-based *computational readability assessment* algorithms that use rich text representations derived from computational linguistics, combined with sophisticated prediction models from machine learning, for deeper, more accurate and robust analysis of text difficulty. These new approaches are dynamic and oriented towards both traditional and non-traditional texts: They can learn to evolve automatically as vocabulary evolves, adapt to individual users or groups, and exploit the growing volume of deep knowledge and semantic resources now becoming available online. In addition, non-traditional domain areas like the Web and social media offers novel challenges and opportunities for new forms of content, serving broad categories of tasks and user populations. This article provides a self-contained survey of automated methods for assessment of text readability: from essential background material, through a summary of current state-of-the-art approaches, to identification of future trends and directions that would benefit from further research.

This survey is intended to complement existing readability-related surveys, which have tended to focus on educational (Benjamin, 2012) or psychological (Zakaluk and Samuels, 1988) aspects of readability measures. The present work provides a computational linguistics and computer science perspective, focusing on core text representations and algorithms used by computational readability assessment methods, and taking a broad view of application areas. Finally, based on the literature survey contained here, as well as the author's expensive experience developing core readability models and applying them in complex application domains like Web search, we identify and discuss specific areas not well covered by existing research. These in turn suggest new directions that we believe are compelling and timely for future research in computational methods for readability assessment.

## 2. Background and Early Research

There is a significant body of work on readability that spans the last 70 years. A comprehensive summary of early readability work may be found in the works of Chall (1958), Klare (1963) and Zakaluk and Samuels (1988). *Traditional readability measures* are those that rely on two main factors: the familiarity of semantic units such as words or phrases, and the complexity of syntax. In order to make these measures straightforward to apply, traditional readability formulas make two major simplifying assumptions. First,

the semantic and syntactic factors are estimated using easy-to-compute proxy variables. For example, a popular proxy variable for a word's semantic difficulty is the number of syllables in the word, and a widely-used proxy variable for a sentence's syntactic difficulty is the sentence's length in words. Second, the ordering of words and sentences is typically ignored: The semantic variables are averaged over all words, and syntactic variables averaged over all sentences, regardless of order. Aspects of reading difficulty associated with higher-level linguistic structures in the text, such as its discourse flow or topical dependencies, are ignored.

The focus on semantic (vocabulary) and syntactic (sentence complexity) features for readability prediction has made sense for many traditional texts. Vocabulary difficulty is known to account for at least 80% of the total variability explained by readability scores for traditional texts, with sentence structure giving a small additional amount of predictive power (Chall, 1958, p. 156-158). Perhaps the most widely-used traditional measure is the Flesch-Kincaid score (Kincaid et al., 1975), which has been implemented as a feature in word processing software such as Microsoft Word™ and is typical of the dozens of similar variants (Mitchell, 1985) that have been developed. The Flesch-Kincaid formula is:

$$RG_{FK} = 0.39 \cdot [AverageWordsPerSentence] + 11.8 \cdot [AverageSyllablesPerWord] - 15.59$$

In general, combining semantic and syntactic features has yielded the best results for traditional settings (Chall and Dale, 1995).

An important sub-class of traditional measures, termed 'vocabulary-based' traditional measures, estimate the semantic difficulty of words in a text by assigning individual words a familiarity or difficulty level based on their occurrence in a pre-specified vocabulary resource. This 'word difficulty' variable then forms the semantic component of the traditional measure, instead of a surface measure such as syllable count. In classic vocabulary-based readability studies, the vocabulary resource is a reference word list that provides information about the familiarity or difficulty of individual words. One widely-known measure of this type is the Revised Dale-Chall formula (Chall and Dale, 1995), which uses the Dale 3000 word list of words familiar to 80% of American fourth-graders. A word is labeled as 'unfamiliar' if it does not occur in the list. The Fry Short Passage measure (Fry, 1990), is also in this family, and uses Dale & O'Rourke's Living Word Vocabulary of 43,000 types (Dale and O'Rourke, 1981) to provide the grade level of individual words in context. In later approaches, the vocabulary resource has been a text corpus: a word's difficulty is defined in terms of its frequency in a large standard collection of representative text. Rarer words with low frequency in the corpus are considered less familiar, and thus, likely to be more difficult, than higher-frequency words. A widely-used measure in this family is the Lexile measure (Lennon & Burdick, 2004: version 1.0), which uses word frequencies from the Carroll-Davies-Richman corpus (Carroll et al., 1971). All of these vocabulary-based measures combine a word unfamiliarity variable to estimate semantic difficulty together with a syntactic variable, such as average sentence length, for estimating sentence difficulty.

While traditional readability formulas like Flesch-Kincaid are widely-available and relatively easy to compute, they also have some serious limitations, especially in the context of the Web and online information access. First, such formulas make strong

assumptions about the text being assessed: They typically assume the text has no noise, or limited noise, and that it consists of well-formed sentences. Second, traditional measures also require significant sample sizes of text, since they become unreliable for passages with less than 300 words (cf. Kidwell et al., 2009). Third, a number of recent studies have demonstrated the unreliability of traditional readability measures for Web pages and other types of non-traditional documents (Si and Callan, 2001; Collins-Thompson and Callan, 2004; Peterson and Ostendorf, 2006; Feng et al., 2009). In general, the reliance of traditional formulas on a small number of summary text features is both a strength and a weakness: Simple formulas are generally easier to implement, but these same formulas have a basic inability to model the semantics of vocabulary usage in context, which becomes important to capture for richer notions of text difficulty.

Finally, traditional readability measures are based only on surface characteristics of text, and ignore deeper levels of text processing known to be important factors in readability, such as cohesion, syntactic ambiguity, rhetorical organization, and propositional density. They also ignore the reader's cognitive aptitudes, such as the reader's prior knowledge and language skills, which are used while they interact with the text. As a result of these limitations, the validity of traditional readability formula predictions of text comprehensibility is often suspect.

In sum, these types of limitations, along with recent opportunities to exploit new computational and data resources, have recently inspired researchers to explore how richer linguistic features combined with machine learning techniques could lead to a new generation of more robust and flexible readability assessment algorithms. We now give background on these developments as they relate to machine learning-based approaches to readability assessment.

## 3. Automated Readability Assessment

The above limitations in traditional formulas, combined with advances in machine learning and computational linguistics, and the increasing availability of training data, helped precipitate a new approach to readability assessment starting in the early- to mid-2000s. François (2009) has called this the 'AI' (Artificial Intelligence) approach to readability. These new approaches typically combine a rich representation of the text being evaluated, based on a variety of linguistic features, with more sophisticated prediction models based on machine learning. Some of these approaches appear similar to traditional readability formulas based on linear regression, in the sense that the parameters of these learning-based approaches are 'fit' to values that minimize prediction error on a corpus of labeled examples. However, unlike traditional methods, advanced machine learning frameworks use dozens or even thousands of features and can express sophisticated 'decision spaces' that are better at capturing the complex interactions between many variables that may characterize document difficulty for different reading levels and readers. In turn, these models often give increased prediction accuracy and reliability for the specific tasks or populations for which they were trained. This section gives an overview of how these learning-based approaches work, and the nature of some representative current implementations.

### *3.1 Readability assessment as a machine learning problem*

As typically defined, a machine-learning approach to readability prediction consists of three steps, as summarized in Figure 1. First, a gold-standard *training corpus* of individual texts is constructed that is representative of the target genre, language, or other aspect of text for which automatic readability assessment is desired. Each text in the training corpus is assigned a 'gold standard' readability level – typically from expert human annotators, but other measures for assigning the label, such as via crowdsourcing, are discussed later. These 'gold standard' labels are proxy estimates of the reading comprehension level for the target population. The standard unit for reading difficulty labels is the grade level, but other scales of measurement are also used. The grade level could be an ordinal value corresponding to discrete ordered difficulty levels, for instance, American grade levels 1 through 12, or it could be a continuous value within a range, to capture within-level gradations, which are especially important for earlier grade levels (e.g. a text at Grade 3.4). Examples of labeled corpora are given in Section 3.4.

**[Insert Figure 1 here]**

Second, a set of *features* is defined that are to be computed from a text. These features capture semantic, syntactic, and other attributes of the text that are salient to the target readability prediction task. As an over-simplified example, a very basic readability prediction model for second-language readers might compute a semantic feature that is the proportion of unfamiliar words in the text relative to an ESL reference list, and a syntactic feature that is the proportion of passive-voice sentences in the text, by using parse trees computed for each sentence. We discuss the types of features used for readability prediction in detail in Section 3.2.

Third, a *machine learning model* learns how to predict the gold standard label for a text from the text's extracted feature values. First, for each training example (i.e., labeled text from the training corpus), the specified features are extracted to form a feature vector that represents the text. Next, the machine learning model is shown these example feature vectors along with their corresponding gold standard labels. The model typically has a set of parameters that control how a text's label is predicted from its feature vector. To train the model, these parameters are adjusted so that the model's label predictions for each text are as close as possible to the corresponding gold standard labels. One commonly-used measure of prediction error is Root Mean Squared Error (RMSE). To find a set of model parameters that is likely to generalize well to new texts, during the training phase, models are typically cross-validated against data unseen by the model. Lastly, the optimized model is applied to a final, previously unseen subset of the gold standard corpus, called the test set, to estimate how well the prediction model is likely to generalize to future texts. This data-driven approach to readability prediction is a very flexible approach to creating or updating a readability measure: It is often easy to retrain the model for different tasks or populations as long as training data are available. We discuss the role of machine learning models in Section 3.3.

Reading difficulty prediction is different from related machine learning tasks like topic prediction or sentiment prediction (Pang & Lee, 2008) that also assign a label or score to a text passage. For readability, the label is arguably more subjective, or at least most user- or population- specific than sentiment detection. In addition, using machine

learning methods that produce models that are easy for humans to interpret can be especially important in readability prediction, particularly for educational applications where teachers or students may need to understand the factors that help explain *why* a text is considered difficult or a good versus poor match for a student.

Because many factors can influence comprehension, assigning a specific readability level to a given text is not an easy task. How hard is this labeling task for people? To our knowledge there have been few readily available published studies of inter-rater reliability for readability labels. There are a number of domain-specific studies, however. For medical information, a study by Ferguson & Maclean (1991) on teacher readability ratings for 60 medical journal articles found low to high inter-rater agreement depending on the dimension of readability being assessed. Those dimensions having the highest inter-rater reliability involved the aspects of readability that were easiest to define and operationalize for the human raters: lexical difficulty, syntactic complexity, and contextual complexity and support (high agreement, Pearson correlation 0.60-0.90) as compared to rhetorical organization (moderate agreement, 0.40-0.60) and information density/topic accessibility (low agreement, 0.00-0.40). For administrative texts, a study by François et al. (2104) found a rather low inter-annotator agreement among experts: the average Krippendorff's alpha, across 7 batches of 15 texts, did not exceed 0.37. In a crowdsourcing setting with a general set of documents, De Clercq et al. (2013) found a Pearson correlation between crowd-based labels and expert labels (at the 'easy' level) of 0.86.

Specific classes of features have been explored for readability assessment that roughly correspond to factors known to affect readability shown in Figure 2.

**[Insert Figure 2 here]**

These broad categories of readability feature types, from 'low' to 'high' level are:

- *Lexico-semantic*: rare, unfamiliar or ambiguous words.
- *Morphological*: rare or more complex morphological particles.
- *Syntax*: grammatical structure.
- *Discourse*:
    - (i) Micro-structural organization of text: use of connectives and other cohesion features to clarify relationships or transitions;
    - (ii) Macro-structural organization: features characterizing explicit, clear argument structure.
- *Higher-level semantics*: use of unusual senses, idioms, or subtle connotation; domain or world knowledge required to comprehend a text.
- *Pragmatic*: contextual or subjective language influenced by genre, e.g. sarcasm.

We discuss studies that have used the more predominant of these feature types in the next sections. Then, we discuss the role of the machine learning models in which these features are used, and the importance of the model versus feature selection in readability prediction effectiveness.

### *3.2 Text features for computational readability assessment*

*Lexico-semantic features*. Reflecting the importance of vocabulary in readability, *lexico-semantic features* capture attributes associated with the difficulty or unfamiliarity of vocabulary, i.e. specific words or phrases in a text. A widely-used feature of lexical difficulty for a word is thus the relative frequency of that word in everyday usage, as measured by its relative frequency in a large representative corpus, or its presence/absence in a reference word list. Several of these semantic word familiarity features were described earlier as the basis for vocabulary-based readability measures. A particular readability prediction model could either use thousands of individual lexical feature values as input (e.g. corresponding to the presence or absence of specific words in a text), or it could form features that are aggregated estimates of lexical difficulty. An example of an aggregated lexical feature is the ratio of unique terms to total terms observed in a text, a statistic known as the *type-token ratio*. The type-token ratio is one of a more general class of *lexical richness* measures that capture the range and diversity of vocabulary in a text (Malvern & Richards, 2012). These statistics capture the tendency for more advanced texts to be authored using a larger vocabulary and exhibit a larger variation in vocabulary than simpler texts of the same length. Other examples of lexical features are shown in Figure 3.

A statistical *language model* is another source of lexical features, and can be thought of as a word histogram giving the relative probability of seeing any given vocabulary word in a text. Statistical language modeling exploits patterns of word use in language. To build a statistical model of text, training examples are used to collect statistics such as word frequency and order. The word lists used in vocabulary-based readability measures like Dale-Chall may be thought of as a simplified language model. The statistical language modeling method described in Collins-Thompson and Callan (2004) greatly generalized this vocabulary-based approach, so that multiple language models are built automatically from training data, typically one for each grade level to be predicted. Such models can capture fine-grained information about vocabulary usage of individual words across levels. Statistical language modeling provides a probability distribution of prediction outcomes across all grade models, not just a single grade prediction. It also provides more data on the relative difficulty of each word in the document. This might allow an application, for example, to provide more accurate vocabulary assistance.

Like the statistical language models of Collins-Thompson & Callan (2004), the Word Maturity measure (Kireyev and Landauer, 2011; Landauer et al., 2011) tracks usage of individual words and phrases as a function of learning stage. However, a key additional ability of the Word Maturity measure is that it accounts for not only how and when a word's frequency changes with learning stage, but how the word's *usage in context* changes, and thus the degree of knowledge a reader is expected to have at any given stage. For example, a word like 'bug' is used in a limited 'insect' sense in early-stage texts, but acquires additional senses and subtleties of meaning, such as 'surveillance device' in more advanced texts. A word's maturity level $f(w, L)$ is a function not only of word $w$ but also a learner level $L$, allowing for the possibility of more personalized readability measures. To model the richness of contexts in which a word appears, the Word Maturity measure uses Latent Semantic Analysis (LSA: Deerwester et al., 1990) to extract the range of typical 'topics' that characterize a word's context at a given

learning stage. To do this, an intermediate corpus is created for each learning stage to be modeled (e.g. each stage might correspond to a grade level). Next, the representative topics of that stage's corpus are computed using LSA. A word at a particular grade or learning stage is represented by a feature vector of LSA topics, which roughly correspond to the spectrum of topics that occur in the contexts where the word is used. A word's feature vector is also computed for a full, adult-stage reference corpus, representing the full range of senses/topics attributed to that word at its most 'mature' learning stage. Finally, the meaning representation of each word (LSA vector) is compared against the corresponding LSA vector of the same word in the most advanced reference model. These difference are aggregated across all words in the text in question, and individual word knowledge is aligned with the measure by adaptive testing over multiple graded texts. Pearson has implemented a beta version (as of this writing) of the Reading Maturity Metric (RMM: http://www.readingmaturity.com/rmm-web/), which includes Word Maturity features as part of a range of computational linguistic features to assess syntactic complexity, coherence, and structural features of the text.

For languages with rich inflectional and derivational morphology that convey meaning e.g. through the choice of different word suffixes and prefixes, such *morphological* features of words can play an important role in assessing readability. For example, Hancke et al. (2012) showed the effectiveness of adding additional morphology features in readability classification for German.

*Psycholinguistics-based lexical features*. Building on earlier research in language acquisition and psycholinguistics, newer automated readability measures have incorporated features that capture cognitive aspects of reading not directly addressed by the surface vocabulary and syntax features of traditional formulas. These types of lexical features include a word's average age-of-acquisition, concreteness, and degree of polysemy. In particular, word concreteness has been shown to be an important aspect of text comprehensibility: previous studies (Paivio et al., 1968; Richardson, 1975) defined concreteness in terms of the psycholinguistic attributes of perceivability (ability to sense an object) and imageability (ability to imagine the object easily and quickly). Tanaka et al. (2013) incorporated these word concreteness attributes into their text comprehensibility measure. Cognitively-based lexical features have been of particular interest in readability measures for second-language learners (Crossley et al., 2008; Vajjala & Meurers, 2012).

*Syntactic features.* Syntactic complexity is known to be associated with longer processing times in comprehension (Gibson, 1998) and is a widely-used factor included in automated readability assessment. The most recent readability prediction methods use a richer set of features to capture a text's syntactic complexity than just the traditional sentence length. It is now typical to use a natural language parser to perform shallow or deep analysis of text, depending on how well-formed the language structure of the target text genre is expected to be. Syntactic readability features are then computed from these parse structures. Figure 3 shows a list of typical syntactic features derived from shallow and deep parsing.

**[Insert Figure 3 here]**

The more advanced syntactic features capture properties of the parse tree that are associated with more complex sentence structure. Pitler & Nenkova (2008) found that of all these syntax-related features they examine, the average number of verb phrases per sentence had the highest Pearson correlation with difficulty (r= 0.42) in their news corpus. In actually training more complex models, the average parse tree depth feature consistently appeared in the best-performing prediction models. Further examples of advanced syntactic features may be found in the studies of Schwarm & Ostendorf (2005), Heilman et al. (2007) and Kate et al. (2010).

*Discourse-based features.* Text is more than a series of random sentences: language exhibits higher-level, longer-range structure by virtue of the dependencies and relationships that exist between its elements. Often, the interpretation of one element in a text may depend on another: this property has been termed *cohesion* (Halliday and Hasan, 1976). At a macro-level, the *coherence* properties of a text, which reflect its logical ordering of arguments and ideas, and systematic organizational structure, can also be considered part of discourse-level structure that affects the readability of text.

Well-organized, cohesive content should be on average more readable than texts that are not, yet properties like cohesion are not captured by traditional readability formulas. Newer automated assessment measures have attempted to remedy this by adding higher-level cohesion- and coherence-related features such as discourse cues, topic continuity from sentence to sentence, idea density, text composition, and logical argumentation. The study by Pitler and Nenkova (2008) was one of the first to explore measures that combined lexical, syntactic, and higher-level discourse features for predicting readability for English texts. Their work empirically demonstrated that discourse relations are strongly associated with perceived text readability and are robust for both predicting and ranking the readability of texts. Recent work has extended the use of higher-level discourse features to other languages, including French (Todirascu et al., 2013; Dascalu, 2014) and Chinese (Sung et al., 2014).

Advances in computational linguistics, starting in the late 1970s, have made it possible to extract a variety of important new higher-level language features from textual material, particularly with regard to cohesion. Coh-Metrix (Graesser and McNamara, 2004) is a computational linguistics tool that has played a prominent role in automated readability assessment, by providing a multi-dimensional set of linguistic and discourse features for text representation. As of version 3.0, Coh-Metrix incorporated 108 different indices (text features), capturing high-level aspects such as:

- degree of referential cohesion (e.g., noun overlap of adjacent sentences)
- deep cohesion (causal events and actions expressed via connectives)
- degree of narrativity (story-telling aspects),
- temporality (degree of consistent tense and aspect).

Coh-Metrix also provides a rich set of standard and cognitively-motivated lexical features, including word concreteness, imagability, and degree of polysemy. Cohesion-type features are being explored for assessing readability in non-traditional genres that do not follow traditional sentence structure. Flor et al. (2013) define a readability measure

for poetry and prose based on what they term *lexical tightness*, which quantifies the fraction of word pairs in a text that are highly related, as estimated by co-occurrence-based or other association measure.

*Higher-level semantic and pragmatic features.* Given that a text is a communication between author and reader, its readability may depend on the reader having some shared domain knowledge or understanding about the world. This may be evident in the use of specific idioms or local references, or more broadly in requiring background knowledge or cultural context. Along another dimension, pragmatic features capture contextual, subjective aspects of meaning that could be of use for readability in maintaining reader motivation and engagement. This might include characterizing the genre of the text (e.g. satire) or the positive/negative sentiment of the text. As one example, Honkela et al. (2012) conducted a study using semantic and pragmatic features derived from topic modeling and sentiment analysis to select stories that were not only relevant to the reader, but also provided emotionally supportive, encouraging content. In general, few computational approaches to readability have tackled the difficult problem of incorporating higher-level semantic and pragmatic features. It is evident that the future potential of computational linguistics and natural language processing to derive features that can reliably capture the highest levels of text difficulty and understanding, such as pragmatics, subtle semantics, and world knowledge, has yet to be fully explored.

### *3.3 Machine learning models for readability prediction*

How are the above features combined to produce a readability prediction using a data-driven machine learning approach (referred to as the *learning framework*)? In most cases, the computational readability measure can be described as a function that maps text to a numerical output value that corresponds to a difficulty or grade level. Depending on the scale of measurement for the output variable, computational readability prediction can be treated as a form of *classification* task (with ordered or unordered category levels), *regression problem* (with continuous-valued levels) or *ranking* problem (with ordered relative levels). In these learning frameworks, the output variable is typically a readability level or score, and the input variables are the set of feature values computed from the text as described above. Most studies cited here are of the regression or classification type. However, some studies, such as those by Pitler & Nenkova (2009) treat readability prediction as a *pairwise preference* learning problem, predicting the relative difficulty of pairs of documents instead of giving an absolute level to each. Extending this idea, Tanaka-Ishii et al. (2010) treated text readability as a ranking problem, combining pairwise assessments of texts to produce an ordering of the texts by reading ease. This is a natural and useful approach for applications that only require a relative ordering, such as the case of a search engine producing a ranked set of results.

Heilman et al. (2008) compared various classification and regression models for readability prediction, including an examination of how the choice of measurement scale affected prediction accuracy. They found that the most effective predictions of reading difficulty resulted from using a proportional-odds prediction model, which assumes an ordinal scale of measurement. In other words, reading difficulty appears to increase steadily as a function of grade level, but not as a linear function. Thus, ordinal regression

models (McCullagh, 1980) are typically a favored choice of learning framework for readability prediction. Various studies have also used learning frameworks such as Gaussian process regression, decision trees and support vector regression (e.g. Kate et al., 2010).

In the end, a compelling question is whether these more sophisticated non-traditional NLP features and machine learning models have improved accuracy over traditional readability formulas. In general, the answer is yes. In one study, François and Miltsakaki (2012) compared the performance of classic and non-classic readability features, using two predictor models: linear regression, and support vector machines. They found that leaving out non-classic predictors hurt prediction performance and that the best prediction performance was obtained using both classic and non-classic features. Depending on the evaluation measure used, support vector machines (Vapnik, 1995) outperformed linear regression in accuracy, but had comparable explanatory power in terms of outcome variability.

Two general conclusions that we can draw after reviewing dozens of studies using machine learning approaches to readability prediction are the following. First, the combination of rich feature representations of text with machine learning frameworks that can exploit them, has proven to be a powerful approach that greatly extends the important foundational research on traditional readability formulas to provide accurate, flexible, and sophisticated computational assessment of readability. Second, in understanding the reasons for improvements of machine learning methods over traditional formulas, we typically find that the nature of the features used as input to the learning framework usually has more effect on performance than the specific choice of learning framework itself. As one example, a representative evaluation was done by Kate et al. (2010), who looked at both the effect of feature choice and learning framework choice. In varying the features, using only lexical features with the best learning framework (bagged decision trees) resulted in a correlation of $r = 0.5760$, using only syntactic features gave $r = 0.7010$, using language model-based features gave $r = 0.7864$, and using all features together gave the highest correlation of $r = 0.8173$. Then, using all features while varying the learning framework, they reported results using Gaussian Process Regression ($r = 0.7562$), Decision Trees (0.7260), Support Vector Regression (0.7915), Linear Regression (0.7984), and Bagged Decision Trees (0.8173). Clearly, the choice of learning framework can matter, but the gains in performance obtainable from changing the learning framework were, for the most part, smaller than gains obtainable from changing the features. In our experience, this is typical of many machine learning studies for readability prediction.

Thus, all things being equal, other considerations beyond basic accuracy may be a dominant factor in selecting a learning framework for readability prediction. For example, it may be important to attach confidence estimates to readability predictions if those predictions are to be used in subsequent tasks like Web search ranking. In such cases, probabilistic learning frameworks like Bayesian regression may be appropriate. In other scenarios, it may be important for users of the measure to understand why a certain prediction was made. Thus, machine learning methods like decision trees (which justify a label prediction in terms of a series of decisions on individual features) or regression models (where the regression weights can be interpreted as importance factors for the features) may be favored.

### *3.4 Evaluation corpora, measures and results*

In this section we address the questions: what evaluation corpora and measures are used to assess the accuracy of readability prediction algorithms? How accurate are current state-of-the-art readability prediction algorithms?

*Evaluation corpora.* The *graded passage* is a basic unit of evaluation in which a paragraph or short story is assigned a grade level or difficulty score, typically by experts at an educational organization or government entity. Traditionally, the main uses of graded passages have been for standardized assessment of reading comprehension, or as part of student reading practice. These same graded passages are often used by researchers to form a corpus for evaluation of readability prediction measures. However, it is very important to understand the process by which the graded passages were created and their grade level determined. Frequently, existing readability measures are used to calibrate graded passages, and so when evaluating new readability measures, there may be a performance bias in favor of those same existing or similar measures that were used to calibrate the passages.

One public resource recently cited in readability evaluations is the collection of texts known as Common Core Appendix B, comprising 168 docs that span levels roughly corresponding to U.S. grade levels 2-12. The passages are tagged by both level and genre, (speech, literature, informative, etc.). Examples are available from http://www.corestandards.org/assets/Appendix_B.pdf

Graded articles for elementary students provided in digital form by the Weekly Reader Corporation (www.weeklyreader.com) for research purposes have been another popular evaluation resource. For example, Feng et al. (2009) used 1433 graded Weekly Reader articles across ages 7-10 as part of their study. Weekly Reader articles in turn have formed part of hybrid collections created by researchers. The *WeeBit* corpus (Vajjala & Meurers, 2012) combines two Web-based text sources (Weekly Reader and BBC Bitesize) that covers five reading levels, with 625 articles per level. The levels map to students in the age range 7-16. Another resource is the 114 articles from Encyclopedia Britannica written in two styles, for adults versus children, originally collected by Barzilay and Elhaded (2003). Similar two-level easy/difficult corpora are available for Wikipedia: simplified English (simple.wikipedia.org) and default English (en.wikipedia.org). A few domain-specific corpora are available, such as the math readability corpus that contains 120 documents labeled on a difficulty scale from 1 to 7 (available at this writing from http://wing.comp.nus.edu.sg/downloads/mwc). In general, many studies have created their own corpora. Access to most of these research corpora can be sought by contacting the authors. For copyright reasons some corpora are restricted from being made freely available. Unfortunately, as of this writing there is still a lack of significantly-sized, freely available, high-quality corpora for computational readability evaluation.

*Evaluation measures.* One widely-used evaluation measure in studies of computational approaches to readability prediction is *rank order correlation* (typically Spearman's *rho*) between the difficulty levels predicted by the readability measure for the reference texts and the 'gold standard' difficulty levels provided for the same reference texts. The advantage of using rank correlation measures for readability evaluation is that only the relative rank ordering of texts is used as the basis for

comparison. There is no need to normalize the readability scores that may be output from the measure, which may be on a very different scale compared to the gold-standard label, or with other measures being compared. Rank correlation measures such as Spearman's *rho* are also robust to outliers, and do not assume an equal interval measurement scale for the reference measures. The Pearson correlation of predicted grade level with gold-standard readability levels is another common evaluation measure.

When the difficulty level is an ordinal variable, some studies have measured prediction accuracy according to the percentage of texts for which the readability measure correctly predicted the correct gold-standard level (rounded to the nearest integer level if the measure produces a real-valued score). While intuitive, this simplistic definition of accuracy ignores the variability of the predictions, i.e. the size of the error made for an incorrect prediction, and thus should not be used as the main evaluation measure. The Root Mean Squared Error (RMSE) is a more robust measure of accuracy used in studies that does penalize algorithms making larger prediction errors compared to the gold-standard level. For machine learning models trained from data, the technique of cross-validation is typically used to assess the likely variability and generalization error. Cross-validation operates by training on different randomly selected subsets of the training data, measuring the prediction error over the remaining test data, and computing the average prediction error over all cross-validation folds.

*Evaluation results*. How accurate are current state-of-the-art readability measures? A recent study by Nelson et al. (2012) assessed the prediction capabilities of six text difficulty measures that included the Lexile measure (MetaMetrics), Degrees of Reading Power (Questar Assessment), and the Pearson Word Reading Maturity Metric. They used five sets of reference texts, which comprised graded passages from various standardized state tests and reading tests, and examples from the American Common Core Standards as well as the MetaMetrics Oasis student reading practice platform. Rank correlations between predicted and actual levels across the six metrics ranged from 0.59 to 0.79 on standardized state passages. Generally, readability measures that used a broader range of linguistic features produced higher correlations than those that just used word difficulty and sentence length features. They also found that metrics tended to make more accurate distinctions among material at lower grades than material at higher grades.

## 4 Applications of Computational Readability Assessment

Perhaps as compelling as new computational approaches to readability prediction are the applications enabled by such prediction methods. For example, tagging Web pages with metadata containing readability estimates enables not only some compelling educational scenarios like grade-appropriate content recommendation, but also some surprising new capabilities like estimating user motivation during Web search, as we describe further below. We now review several important extensions and applications of automated readability prediction that have been developed for different tasks and populations.

*Readability for Second-Language Learners*

First-language (L1) readers have very different skills and needs compared to second-language (L2) readers. A key difference between L1 and L2 readers is the timeline and processes by which language are acquired. For L1 learners, acquisition starts in infancy, and primary grammatical structures are typically acquired by age four (Bates, 2003) – prior to the start of the child's formal education. L2 readers are often college-age or older, have a sophisticated conceptual lexicon, and can grasp complex ideas and arguments. Second-language learners, on the other hand, unlike their L1 counterparts, are still actively involved in learning the grammar of the target language, so even intermediate and advanced students of second languages, who correspond to higher L2 readability skills, can struggle with grammar in the target language.

While most development of readability measures has focused on L1 readers, a number of recent studies have developed automated readability assessment methods that try to account for these special aspects of second-language L2 learners. One of the first studies to develop machine learning-based readability measures for L2 readers was that of Heilman et al. (2007), who showed that grammatical features may play a more important role in second-language readability prediction than in first-language readability. Other automated measures for English readability for L2 readers subsequently were explored in work by Crossley et al. (2008), who used a rich feature set computed by the Coh-Metrix computational tool that included syntactic sentence similarity, lexical co-referentiality, and word frequency. Schwarm and Ostendorf's work (2005) on general readability prediction was partly motivated by the need for tools in bilingual education.

*International Language Support*

In the past, the majority of traditional readability assessment research focused on English, with other languages adapting and extending those results. For example, after the Flesch formula for readability of English text (Flesch, 1948) was published, a series of adaptations for European and other languages followed: Kandel and Moles (1958) published an adaptation for French, and soon after, José Fernández Huerta (1959) published a corresponding formula for Spanish text that is still widely used. Zakaluk & Samuel (1988) contains a comprehensive list of traditional readability formulas for a wide variety of languages. More recently, much original research has originated with languages other than English, and in cases where English studies are published early on, adaption to other languages is happening on a more compressed time scale than happened with traditional methods. In particular, Asian and European languages have become early originators and adapters of improved computational methods.

Varying degrees of effort are needed to re-purpose machine learning-based automated assessment methods originally developed for one language (e.g. English) to other languages. This effort depends on factors such as the linguistic complexity of the features required by the automated method, the existence of a gold-standard training corpus in the target language of appropriate quality and size, and the linguistic nature of the target language itself. Computing linguistically complex features, such as syntactic difficulty features derived from parse trees, requires natural language processing (NLP) tools such as parsers trained for the target language, which may not be available. The lack of adequate training corpora in some target languages has arguably been a bottleneck

in deploying automated processes for a variety of NLP-related tasks, from parsing to readability assessment. Finally, knowledge of the nature of the target language will influence the type of feature extraction required for readability assessment. For example, for highly inflected languages like French or Russian, morphology becomes critical to consider as part of computing semantic difficulty features. There are also specialized features that are unique to some classes of languages. For example, Chinese readability formulas include features based on character symmetry and number of strokes (Lau, 2006).

Among those recently applying new semantic resources and learning-based computational methods are languages as diverse as Chinese (Lau, 2006; Chen et al., 2013), German (Vor Der Brück and Hartrumpf. 2007), French (François and Fairon, 2009), Arabic (Al-Khalifa and Al-Ajlan, 2010), Japanese (Tanaka-Ishii et al., 2010), Thai (Daowadung and Chen, 2011) and Swedish (Sjöholm, 2012). Due to the aforementioned lack of multi-level graded corpora for languages other than English, researchers have built readability models from freely available collections of two or three classes collected from the Web. Dell'Orletta et al. (2011), Aluisio et al. (2010), and Klerke and Søgaard (2012) report on creating and experimenting with such corpora in Italian, Portuguese and Danish respectively.

*Supporting Readers with Disabilities*

In addition to native and non-native speakers from different locales, readability measures are starting to be adapted for those with language learning disabilities and dyslexia. Abedi et al. (2003) examined classic readability features for reading test items in order to identify those grammatical and cognitive features that differentially contribute to reading difficulty for students with disabilities, and thus have a negative impact on performance. Their study focused on Grade 8 students and reading assessments, and thus further research would be required to understand if and how their findings generalize. However, within this population they found that certain surface textual/visual features had the highest discriminative power between students with and without disabilities, such as the use of long words (greater than seven letters in length), suggesting that changes in font, word length and spacing, and reduction in distracting visuals were important factors in readability for that target population. Related findings were made by Rello et al. (2013) for readers with dyslexia: Comprehension was independent of readability, and word length was critical, with shorter words helping comprehension. Sitbon & Bellot (2008) developed a sentence readability measure for dyslexic readers based on features informed by traditional readability measures (i.e. the French version of the Reading Ease score) as well as psycholinguistic studies on the reading processes of dyslexic readers (predicted reading time based on phoneme cohesion, number of adverbs and conjunctions). Feng et al. (2009) developed and evaluated automated readability assessment tools for readers with intellectual disabilities, exploring the use of cognitively-motivated features such as the 'entity density' – the number of entities mentioned per sentence. They reported higher Pearson correlation with comprehension scores (for adults with intellectual disabilities) for readability models trained with cognitively-motived features, compared to standard lexical and syntactic features. Beyond assessment, techniques for text simplification and

summarization hold promise as approaches to improving readability for learners with special needs, such as dyslexic learners (Nandhini & Balasundaram, 2011).

*Computer-assisted Educational Learning Systems*

Many educational scenarios require the ability to find information at the right level of difficulty, or of the right type of difficulty, for a student. Thus, automated readability measures can play a central role in educational settings, particularly for language learning and reading tutoring systems. For example, an online language tutor might find authentic examples of high-quality Web content that were tailored to individual student goals in order to help them learn new vocabulary in realistic contexts. Like people, intelligent systems would need an ability to find relevant material at the right level of difficulty, quickly and precisely. Unlike people, an application might use long, complex queries that expressed multiple specific constraints that good pages should satisfy: using the right target vocabulary, at the right level of difficulty, without too many other unknown words, and so on.

One example of such a system is the REAP vocabulary tutor developed at the Language Technologies Institute of Carnegie Mellon University (http://reap.cs.cmu.edu). REAP uses sophisticated filtering and ranking technology to deliver personalized language instruction in English, French, and Portuguese. REAP has helped hundreds of second-language learners in classrooms, while also providing a fascinating experimental platform to study what helps students learn vocabulary most effectively. In one controlled study (Heilman et al., 2010), using REAP's ability to personalize examples to individual students' self-reported topical interests led to consistent gains in student performance in vocabulary acquisition, compared to a control group on the same system without personalization.

In related work, Beinborn et al. (2012) study the applicability of readability measures to self-directed language learning, and argue for assessment over individual dimensions of readability (as in Figure 2) rather than overall readability predictions, in addition to modeling the background knowledge of the learner. We also note the development of classroom-oriented tools like ReaderBench (Dascalu, 2014), an environment for analyzing text complexity and reading strategies that explicitly incorporates rich text representation, including advanced readability features capturing discourse structure.

*Readability Prediction for the Web*

The highly varied, non-traditional nature of Web content, from blog comments to search engine result pages to online advertising, leads to new challenges for readability prediction. In addition to text with non-traditional structure, Web pages can also contain images, video, audio, tables, and other rich layout elements that can influence text readability. The ability of a user to understand a document would seem to be a critical aspect of that document's value, and yet a document's reading difficulty is a factor that has typically been ignored in designing access to Web content.

This lack of attention to readability has been especially true for Web search engines – one of the primary ways people access information on the Internet. While some

work (e.g. Kanungo and Orr, 2009) has recognized the importance of readability as a crucial presentation attribute of search results and other Web summaries, traditionally search engines have ignored the reading difficulty of documents and reading proficiency of users as part of their retrieval process. For domains like online health care resources for the elderly (Becker, 2004) and educational resources for children and students (Collins-Thompson et al., 2011), there is a need not only for more accessible content, but for better ways to find such content if it already exists. While providing accessible content via a search engine requires solving many important problems in interface design, content filtering, and results presentation, one fundamental problem is simply that of providing relevant results at the right level of reading difficulty.

An initial step in solving this problem is to label existing Web pages with metadata that contains readability estimates. Beyond its utility for basic Web search, enriching Web pages with readability metadata has led to a variety of new and sometimes surprising applications (Collins-Thompson, 2013). Figure 4 summarizes the impact that readability metadata can have in enabling new capabilities for information systems of the Web. For example, there is a natural connection with the problem of modeling user and site expertise (Kim et al., 2012).

**[Insert Figure 4 here]**

Unlike traditional texts, Web pages have valuable additional sources of information by virtue of their hypertext representation, such as the set of links to and from the page, and the anchor text associated with those links. This additional context has been used to improve readability estimation for individual pages and predict the appropriateness of pages for children. Gyllstrom and Moens (2010) proposed AgeRank, an algorithm that provides a binary labeling of Web documents according to its appropriateness for children versus adults. The page's age-appropriateness label is inferred using a graph walk algorithm inspired by the PageRank algorithm that Google introduced to estimate the importance of Web pages. The AgeRank approach also uses features such as page color and font size to help determine the page label. The combination of Web graph, vocabulary, and non-vocabulary features with existing machine learning methods is likely to provide a good basis for estimating the readability of Web documents. In related work, Akamatsu et al. (2011) proposed a method to predict the comprehensibility of web pages that uses hyperlink information in addition to textual features. The authors showed reasonably high positive correlation between the link structure and readability levels of pages on the Web.

In general, little is currently known about basic readability properties of the Web, or the influence of readability on user interactions with Web content. Thus, there is a need for large-scale reading-level analysis of the Web that examines properties like the relationship of reading level metadata to other meta-data for the same pages, such as a page's topic, analysis of differences in reading level distributions across different domains and types of pages, such as high- versus low-traffic pages, and interesting hyperlink-based clusters with low and high inter-page differences in reading level. Some recent work has begun to study Web readability via user interactions captured in search engine query logs. Duarte Torres et al. (2011) performed an analysis of the AOL query log to characterize so-called 'Kids' queries. A query was labeled as a Kids query if and

only if it had a corresponding clicked document whose domain was listed as an entry in the 'Kids&Teens' category of the Open Directory Project. More analysis is needed to obtain a better understanding of where and how readability meta-data is likely to be most effective for specific search tasks or groups of users on the Web.

To match users to Web content, a search engine or recommendation algorithm needs to represent and estimate the reading proficiency of the user. Children may not want material that is too difficult, and experts may want highly technical content, not tutorials and introductory texts. Non-native language speakers also form a significant population of users who could benefit from search engines that can account for the reading level of both users and content. One approach to representing reading proficiency is to have users self-identify their level of proficiency or desired material. This is the approach Google has used in their deployment of an Advanced Search feature to filter results by Low, Medium, and High levels of difficulty (Russell, 2011). However, self-identified user information may not always be available or reliable, in which case we need ways to construct a reading proficiency profile automatically. Initial work on automatically estimating a reading proficiency profile for a specific user from their interaction with a Web search engine was introduced by Collins-Thompson et al. (2011) and by Tan et al. (2012). In future work, we expect that existing learning algorithms could be applied to learn user readability profiles based on observations such as the reading level of pages that were recently read; semantic or syntactic features of current and past queries; previously visited pages or domains from a known list of expert or kids-related sites; and other features of the user's history or behavior. More generally, we foresee the need for topic-specific models of readability that reflect a user's expertise on specific topics but not others, in addition to their general reading proficiency.

Web search engines whose results rankings account for user and content readability levels aim to reduce the 'gap' between the user's estimated reading proficiency profile and a document's reading difficulty profile. As with other types of personalization, there is a risk-reward tradeoff: We want to promote easy-to-read documents closer to the user's reading proficiency level, while not straying too far from the default ranking, which is typically a highly-tuned relevance signal optimized for the 'average' user. Moreover, we may want to show the user pages that 'stretch' their reading ability in order to help them learn about a new topic.

Research in applying meta-data derived from reading level prediction to the Web and other information retrieval domains is only just beginning. We believe it has the potential to improve the performance of a wide range of online tasks for individual users, from personalized Web search to educational applications.

## 5 Classification of Existing Computational Readability Approaches

Before describing future avenues of research, it is worth taking a high-level view of recent readability literature to find opportunities to improve the coverage of existing research. Figure 5 provides a visual classification of representative papers covered in this article that have introduced new automated readability assessment methods, most within the past decade, for different tasks or target populations. Papers (identified with a short citation key) have been classified in the horizontal direction according to the primary type or combination of features used to predict readability, and in the vertical direction

by primary population or task. Some of the papers span multiple features or target populations. We regret that many interesting papers had to be excluded from this overview in the interest of clarity and space. This overview is focused on features of text and does not include, for example, readability prediction using behavioral cues such as eye movements that are not (yet) widely used for computational assessment.

**[Insert Figure 5 here]**

This visual summary of the automated assessment landscape reveals several areas where current research is lacking. First, in general, a limited number of readability models have incorporated features of higher-level text structure. This is particularly the case for languages other than English, most likely due to the current sparsity of linguistic resources and tools needed to estimate and assess models for those locales. Second, the same may be said about specialized domains that include technical or scientific writing and poetry/prose. For example, initial versions of readability measures have been developed for health informatics (e.g. Wang 2006) that typically focus on word familiarity. There is also a lack of prediction approaches for more fine-grained readability prediction, such as at the sentence level – although researchers have begun exploring this area (Pilán et al., 2014), particularly in the context of text simplification (Vajjala & Meurers, this issue). In addition, only a few studies have incorporated pragmatic, genre-related features for readability assessment in any language. Third, there has been little published work on automatically learning personalized models of individual reading expertise. The closest relevant work so far published in that area (e.g., Collins-Thompson et al., 2011) has been related to Web search, where data from query logs has made it possible to estimate anonymous, personalized models of user interests and expertise by using behavioral signals such as queries issued and documents clicked. These models, in turn, have been used to improve the quality of Web search ranking for individual users.

## 6 Future Research Directions

Based on the state of existing research summarized above and trends in increasing availability of data and computing resources, in this section we propose three complementary directions in which future research on computational approaches to readability modeling and prediction is needed. We then discuss several specific research directions in more detail.

1. **User-centric models**. Text readability has an inherently individual, subjective component that current readability measures do not adequately capture. However, developing personalized and adaptive measures will require new approaches to evaluation and validation: The usual gold-standard approach for assigning readability labels is no longer appropriate since generic labels may not reflect an individual user's context or knowledge. Moreover, users are dynamic individuals whose expertise and interests evolve over time, and who may use different styles of learning and strategies for overcoming comprehension difficulties. Adapting users to content (personalized training)

and adapting content to users (personalized simplification) are two potential research directions mentioned below.

2. **Data-driven measures**. Machine learning models require data for training – which is both their strength and weakness. To obtain labeled data, the use of human computation and crowdsourcing are promising avenues that researchers are beginning to explore. Readability measures tailored for new types of data, such as new content formats like blogs, wikis, online surveys, and writing genres, will continue to play an important role in Web interaction, especially for educational settings. The dynamic nature of the Web and constant introduction of new vocabulary into the world's languages also mean that effective readability measures will need to continuously evolve to reflect these changes. Further data-driven readability measures are needed that are easily specialized for specific domains, like health care or scientific content, using methods that do not require an external corpus or hand-graded labels. One attempt in that direction was the unsupervised model of Jameel et al. (2012). They proposed an initial model that computed technical readability, making two assumptions: first, that documents containing rarer terms deviating from the common terms would be more technically difficult, and second, the more cohesive the terms (words or short phrases) within a text, the more technically simple the text. Further work in this area is needed.
3. **Knowledge-based models**: To achieve deeper content understanding for readability prediction will require corresponding advances in natural language tools and machine learning frameworks – including projects that attempt to model world knowledge. One specific research challenge requiring such broader knowledge is to identify *gaps* and *unstated assumptions* that are a higher-level source of difficulty. Another example is that while we have the ability to model the topics that are discussed in text, little work has been done on capturing the *dependencies* between concepts. That is, algorithms that can 'understand' what a user needs to know *before* they can understand a second concept will prove invaluable sources of assistance for many tasks. Health informatics is one important application area that would benefit from these types of advances that capture and exploit deep domain knowledge.

Making progress in these directions will require a combination of new approaches and resources. In particular, key aspects of further progress that need to be developed or encouraged in the computational linguistics and computer science communities include the following.

1. *Improved annotated data resources.* One challenge to the advancement of automated readability research has been the lack of representative corpora and associated datasets, especially for languages other than English. The issue of digital copyright has been one factor in the difficulty of sharing resources. New, freely available corpora need to be developed that would encompass a broad variety of text genres, media types, and document properties (from longer full texts, to short text snippets) with difficulty labels from many human assessors. When constructed properly, such resources would provide

a basis for training data-driven methods, designing reproducible experiments, evaluation of corpus-based methods, and objective comparison across algorithms. The advent of crowdsourcing is likely to help with label acquisition, as discussed later in this section.

2. *Standardized, realistic task definitions and evaluation methodology* to be applied with the above datasets: It is typical of many papers introducing automated text difficulty assessment that they often report results solely in terms of correlations with other existing automated measures, without checking their effectiveness on a real-world task or desired outcome. A more organized effort in the computational linguistics community that standardizes task and evaluation criteria, like those already organized annually for other tasks such as summarization, entity-finding, and information retrieval, would rapidly advance the cause of automated readability assessment.

3. *Inter-disciplinary collaborations*. The problems associated with understanding and modeling text difficulty for individual readers are inherently multi-disciplinary. Research progress will depend on paradigms and methods spanning linguistics, education, psychology, computer science and other fields. Thus, community-building activities such as workshops that build cooperation across fields would help to fully develop the potential of computational readability methods.

We now give more detail on a few potential research directions that reflect these goals.

*Adaptive and personalized readability algorithms*

Instead of assuming the reading level of users and documents is something to be passively observed, a new class of algorithms that we term *adaptive readability* algorithms could seek optimal strategies and methods for augmenting content or user knowledge in order to actively reduce the 'knowledge gap' between the author and a particular reader. For example, when recommending a Web site to a user whose difficulty is higher than the user's current proficiency, an adaptive readability system could perform personalized user training, first identifying important *words to learn* on the site's pages that the user is not likely to know – e.g., an article about stomach aches might use the technical term gastritis. The system could provide links to supporting definitions, background material, or a simplified version of the text that uses the more familiar words.

Such adaptive algorithms would need to be able to solve problems that include: enabling personalized readability estimation by computing and maintaining a dynamic reading proficiency and domain knowledge model for each user; identifying key vocabulary in a document; comparing this key vocabulary against the user's reading proficiency model; and computing the best small subset of critical 'stretch' vocabulary required to understand most of a document. Other relevant scenarios include intelligent tutoring applications that help stretch the student's vocabulary by retrieving content that is slightly above their current reading level, along with satisfying other linguistic properties that align with curriculum goals. The REAP intelligent tutor (Collins-

Thompson and Callan, 2004) mentioned earlier is an example of a first step toward this type of functionality. In a related direction, Agrawal et al. (2011) used estimates of syntactic complexity and key concepts to identify difficult sections of textbooks that could benefit from better exposition and to find links to authoritative content. Algorithms for automatic text simplification (Siddarthan, 2014) could play a highly complementary role to readability measures, producing summarizations with personalized knowledge of which words a user knows or doesn't know based on their reading proficiency profile. The educational potential for such augmentations, especially those that are personalized based on individual user models, seems very compelling.

*Local readability estimation*

Any given text might display large variations in difficulty across different sections of the document. This is especially true for longer texts commonly occurring across a variety of genres, including book chapters, movie scripts, legislative texts, and product documentation. As compared to the 'global' difficulty estimate for the entire document, 'local' variations in difficulty can come from a number of factors, including changes in topic, quotations of external material, change of character in dialogue, and so on. Previous readability studies have explicitly acknowledged this phenomenon by prescribing application procedures that sample passages throughout a text, and then combining the readability levels of the sampled passages to produce an overall readability score.

Kidwell et al. (2009, 2011) introduced an explicit local readability estimation approach that applied a locally weighted version of a global readability model to a sliding window of width *k* words (e.g., 100 words). As the window moved from the beginning of the document to the end, a sequence of readability scores was generated, one per window. The degree of locality was controlled with the width parameter *k*, which could also be viewed as controlling the degree of smoothing of readability estimates. A narrow window emphasized scrutiny of local behavior, such as a specific paragraph or conversation.

Describing local readability variation as an object of study is valuable in itself, especially when combined with visualization methods. Local estimation will enable future applications as diverse as improved document summarization, identifying interesting events in a long text or transcript, and finding difficulty 'hotspots' in textbooks or documentation that need additional simplification, explanation or augmentation.

*Real-time readability assessment from behavioral signals*

The advent of inexpensive, increasingly accurate sensor equipment and analysis software for tracking human behavioral signals, such as eye movement and electrical brain activity, provides a promising new source of cues about text difficulty that could be integrated as features in prediction settings, especially in real time. Ultimately, such signals could assist in estimating individual cognitive difficulty or ease at both the decoding level and higher cognitive levels. A recent study by Cole et al. (2012) showed that a user's level of domain knowledge could be estimated from real-time measurements

of eye movement patterns during search tasks. Researchers have also begun exploring non-invasive assessment of reading comprehension using low-cost EEG detectors that monitor electrical activity in the brain via detectors on the surface of the scalp. In one early study, Chang et al. (2013) found that some EEG signal components appear to be sensitive to certain lexical features. For example, they found a strong relationship between a word's age-of-acquisition, and activity in the 30-100Hz EEG frequency band for child subjects, along with a number of weaker correlations with other lexical features like word frequency in adult subjects. While many technical difficulties remain in accurately estimating mental states and activity from such behavioral signals, their potential to contribute to our understanding of reader engagement and comprehension is a promising avenue for future automated readability assessment methods.

*Crowdsourcing for readability annotation*

Traditionally, graded passages that serve as learning materials and training examples for machine learning approaches to readability have been developed by experts. Thus, one significant issue in data-driven reading difficulty modeling and prediction has been that it is time-consuming and expensive to obtain the needed difficulty labels manually assigned by experts. This is one reason for a subsequent lack of labeled corpora, and a large number of expert-labeled examples are typically needed by the learning framework to fit the parameters of the readability models.

The rise of crowdsourcing platforms such as Amazon Mechanical Turk (AMT), however, have made it possible to gather readability judgments from a large, diverse audience of non-experts that, in aggregate, have the potential to approach expert quality at a fraction of the cost. A crowdsourcing platform like AMT or Crowdflower is typically a Web-based service that serves as a marketplace connecting people willing to complete online tasks for pay (crowd workers) with those needed the online tasks completed with good accuracy (task authors). Tasks that are a good fit for crowdsourcing are those that are easy for human intelligence, but difficult for machine intelligence. The assessment of a complex phenomenon like text difficulty certainly qualifies as a good fit. Typically, the quality of the non-expert crowdsourcing labels is maintained through the use of mechanisms such as randomly inserted assessment tasks using a small number of known, expert-labeled answers which the crowd worker must answer at a high level of accuracy in order to be fully compensated for their work. As an example of cost, to obtain more than 5,000 reliable pair-judgments over several hundred passages (Chen et al., 2013) cost on the order of US$250, or about 5 cents per pair.

One of the first studies to examine the use of crowdsourcing to obtain readability assessments was that of De Clercq et al. (2013). Their study used expert readers to rank texts according to relative difficulty. They compared these expert rankings to rankings derived from the use of a crowdsourcing tool where non-expert users provided pairwise comparisons about the relative difficulty of two texts. The non-expert labels were of comparable quality to the expert labels. Independently, Chen et al. (2013) developed an efficient statistical model to combine the pairwise assessments from a budgeted number of crowd workers into an aggregate ranking of reading difficulty. Their study introduced an active learning method that was shown to reduce the cost (in terms of the number of

non-expert crowd assessments) required to achieve a given level of ranking accuracy compared to a reference ranking generated from expert labels.

Given the volume and variety of labelled data that will be required to drive the retraining of future machine learning-based methods for different tasks, domains, and target populations, algorithms that optimize efficient crowdsourcing of readability labels or features are likely to be a fruitful tool, and an on-going research in their own right. However, while crowdsourcing appears to show promise as a source of readability annotations, several caveats are also in order. First, the quality and nature of results obtained from crowdsourcing can be very sensitive to details in the task and interface design (Kittur et al., 2008). Second, readability assessment is highly dependent on the reader's profile, and thus may suffer in generic crowdsourcing scenarios. Third, some researchers have raised ethical and legal issues, such as potential worker exploitation, in the use of crowdsourcing platforms (e.g. Fort et al., 2011). Beyond crowdsourcing, other avenues such as 'games with a purpose' (von Ahn & Dabbish, 2008) that could generate annotation data or solve related computational readability problems as an outcome of game play, may serve as fruitful alternatives to explore in future research.

## 7 Conclusion

Computational methods for readability assessment promise to provide a powerful technological tool that will touch many aspects of how we interact with, learn from, and discover information. While the nature of texts and readers will continue to evolve, the basic need for algorithmic methods that model and estimate text difficulty and readability is as strong as ever. The past ten years have seen a fundamental shift in approach: from traditional general-purpose formulas with two or three variables that are fitted with small amounts of expert-labelled data, to machine-learning based frameworks that use a rich feature representation of documents trained from large corpora using aggregated, non-expert crowdsourced labels, along multiple dimensions of representation that capture deeper aspects of text understanding and difficulty.

Our review of the field highlighted the lack of published research in areas such as data-driven and personalized readability measures, and test collections and evaluation measures for non-traditional texts. We believe this is due to two factors: the novelty of the field, and the methodological and technical difficulties in developing and evaluating reliable personalized models. Future challenges include balancing the relevance and comprehensibility of texts, and richer document representations for enhancing readability dimensions. The next ten years will bring further developments in personalized, data-driven, deep knowledge-based models of text readability. It seems likely that statistical machine learning will play a key role in future development of readability measures, providing a principled framework that can learn from data and handle the rich sets of complex features and decision spaces that are required to capture deeper text understanding.

Computational text readability assessment continues to be a promising field that tackles problems at the heart of human language understanding. The need for automated assessment of text readability will exist as long as there is human language and the desire for people to learn and inform each other, and as long as our computational models of language and language acquisition continue to grow. User-centric, data-driven,

knowledge-based text readability assessment is an exciting and promising research direction that connects deeply with our most difficult research problems in modeling and interpreting human language. Advances in text readability assessment will act as a key that unlocks a rich array of applications that help people learn and communicate, whether in elementary school or for a lifetime.

---

**References**


Abedi, J., Leon, S., Kao, J., Bayley, R., Ewers, N., Herman, J., & Mundhenk, K. (2011). Accessible Reading Assessments for Students with Disabilities: The Role of Cognitive, Grammatical, Lexical, and Textual/Visual Features. CRESST Report #785. Univ. of California, Los Angeles. Jan 2011. http://www.cse.ucla.edu/products/reports/R785.pdf

Agrawal, R., Gollapudi, S., Kannan, A., & Kenthapadi, K. (2011). Identifying enrichment candidates in textbooks. In *Proceedings of the 20th International Conference on World Wide Web (WWW '11)*. ACM, New York, NY, USA, 483–492.

Akamatsu, K., Pattanasri, N., Jatowt, A., & Tanaka, K. (2011). Measuring Comprehensibility of Web Pages Based on Link Analysis. In *Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, vol. 1, 40-46.

Al-Khalifa, H. S. & Al-Ajlan, A. A. (2010). Automatic Readability measurements of the Arabic text: An exploratory study. *Arabian Journal for Science and Engineering*, 35.

Barzilay, R. & Elhadad, N., (2003). Sentence Alignment for Monolingual Comparable Corpora, In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP'03)*, 25-32.

Bates, E. (2003). On the nature and nurture of language. In R. Levi-Montalcini, D. Baltimore, R. Dulbecco, & F. Jacob (Series Eds.) & E. Bizzi, P. Calissano, & V. Volterra (Vol. Eds.), Frontiers of biology. The brain of homo sapiens. Rome: Istituto della Enciclopedia Italiana fondata da Giovanni Trecanni S.p.A., pp. 241-265.

Becker, S.A. (2004). A study of web usability for older adults seeking online health resources. *ACM Transactions on Computer-Human Interaction (TOCHI)* 11, 4, 387–406.

Beinborn, L., Zesch, T., & Gurevych, I. (2012). Towards fine-grained readability measures for self-directed language learning. *Proc. of the SLTC 2012 workshop on NLP for CALL: Linkoping Electronic Conf. Proceedings* 80: 11-19.

Benjamin, R. (2012). Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. *Educational Psychology Review*, 24(1):63-88.

Carroll, J. B., Davies, P., & Richman, B. (1971). Word Frequency Book. Boston: Houghton Mifflin.

Chall, J.S. (1958). Readability: An appraisal of research and application. Bureau of Educational Research Monographs, No. 34. Columbus, Ohio State Univ. Press.

Chall, J.S. & Dale, E. (1995). Readability Revisited: The New Dale-Chall Readability Formula. Cambridge, MA: Brookline Books.

Chang, K.M., Nelson, J., Pant, U., & Mostow, J. (2013). Toward Exploiting EEG Input in a Reading Tutor. *International Journal of Artificial Intelligence in Education*, 22 (1-2), 19-38.

Chen, X., Bennett, P.N., Collins-Thompson, K., & Horvitz, E. (2013). Pairwise ranking aggregation in a crowdsourced setting. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining (WSDM '13)*. ACM, New York, NY, USA, 193–202.

Chen, Y.-T.; Chen, Y.-H. & Cheng, Y.-C. (2013). Assessing Chinese Readability using Term Frequency and Lexical Chains. *Computational Linguistics and Chinese Language Processing*, 18, 1-18.

Cole, M.J., Gwizdka, J., Liu, C., Belkin, N.J., & Zhang, X. (2012). Inferring user knowledge level from eye movement patterns. *Information Processing and Management*. DOI: http://dx.doi.org/10.1016/j.ipm.2012.08.004

Collins-Thompson, K., Bennett, P.N., White, R.W., de la Chica, S., & Sontag, D. (2011). Personalizing web search results by reading level. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management (CIKM '11)*. ACM, New York, NY, USA, 403–412.

Collins-Thompson, K., & Callan, J. (2004b). 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 (SIGIR '04)*. ACM, New York, NY, USA, 544–545.

Collins-Thompson, K, & Callan, J. (2004c). A Language Modeling Approach to Predicting Reading Difficulty. In *Proceedings of HLT-NAACL 2004*. 193–200.

Collins-Thompson, K. & Callan, J. (2005). Predicting reading difficulty with statistical language models; *Journal of the American Society for Information Science and Technology*, 56, 1448-1462.

Collins-Thompson, K. (2013). Enriching the web by modeling reading difficulty. In *Proceedings of the Sixth International Workshop on Exploiting Semantic Annotations in Information Retrieval (ESAIR '13)*. ACM, New York, NY, USA, 3-4.

Crossley, S. A., Greenfield, J., & McNamara, D. S. (2008). Assessing text readability using cognitively based indices. *TESOL Quarterly*, 42 (3), 475-493.

Dale, E., & Chall, J.S. (1949). The concept of readability. *Consciousness and Cognition* 26(23) (1949).

Dale, E., & O'Rourke, J. (1981). The Living Word Vocabulary. Chicago, IL: World Book/Childcraft International.

Daowadung, P., & Chen, Y.-H. (2011). Using word segmentation and SVM to assess readability of Thai text for primary school students. *International Joint Conference on Computer Science and Software Engineering: JCSSE*, 2011.

Dascalu, M. (2014). ReaderBench (2)-Individual Assessment through Reading Strategies and Textual Complexity. *In Analyzing Discourse and Text Complexity for Learning and Collaborating*. Springer International Publishing. 161-188.

De Clercq, O., Hoste, V., Desmet, B., van Oosten, P., De Cock, M., & Macken, L. (2013). Using the Crowd for Readability Prediction. *Natural Language Engineering*. 1(1). Cambridge University Press.

Deerwester, S., Dumais, S.T. Furnas, G.W., Landauer, T.K., Harshman, R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science* 41 (6): 391–407.

Duarte Torres, S., & Weber, I. (2011). What and how children search on the web. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management (CIKM 2011)*, 393-402.

Eickhoff, C., Serdyukov, P., & de Vries, A.P. (2011b). 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*. ACM, 505–514.

Feng, L., Elhadad, N., & Huenerfauth, M. (2009). Cognitively motivated features for readability assessment. In *The 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2009).*

Ferguson, G., & Maclean, J. (1991). Assessing the readability of medical journal articles: an analysis of teacher judgements. *Edinburgh Working Papers in Linguistics*. No. 2, 112-125. http://files.eric.ed.gov/fulltext/ED353790.pdf

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, Vol 32(3), Jun 1948, 221-233.

Flor, M., Klebanov, B. B., & Sheehan, K. M. (2013). Lexical Tightness and Text Complexity. *Proceedings of the Second Workshop on Natural Language Processing for Improving Textual Accessibility*, 2013.

François, T. L. (2009). Combining a statistical language model with logistic regression to predict the lexical and syntactic difficulty of texts for FFL. *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, Association for Computational Linguistics.

François, T. & Fairon, C. (2012). An "AI readability" formula for French as a foreign language. In *Proceedings of the 2012 Conference on Empirical Methods in Natural Language Processing (EMNLP 2012)*, 466-477.

François, T. & Miltsakaki, E. (2012). Do NLP and machine learning improve traditional readability formulas? *Proceedings of the First Workshop on Predicting and Improving Text Readability for target reader populations*, Association for Computational Linguistics, 2012, 49-57.

François, T., Brouwers, L., Naets, H., & Fairon, C. (2014). AMesure: une formule de lisibilité pour les textes administratifs. In *Actes de la 21e Conférence sur le Traitement automatique des Langues Naturelles (TALN 2014)*, Marseille, 467-472.

Fort, K., Adda, G., & Cohen, K.B. (2011). Amazon Mechanical Turk: Gold Mine or Coal Mine? Last Words editorial. *Computational Linguistics* 37:2.

Fry, E. (1990). A readability formula for short passages. *J. of Reading*, May 1990, 594-597.

Gibson, E. (1998) Linguistic complexity: locality of syntactic dependencies. *Cognition*, 68:1-76.

Graesser, A.C., McNamara, D.S., Louwerse, M.M., & Cai, Z. (2004). Coh-metrix: analysis of text on cohesion and language. *Behavior Research Methods, Instruments, and Computers* 36(2) (2004), 193–202.

Gyllstrom K., & Moens, M-F. (2010). Wisdom of the ages: toward delivering the children's web with the link-based agerank algorithm. In *Proceedings of the 19th ACM international conference on Information and knowledge management (CIKM '10).* ACM, New York, NY, USA, 159–168.

Halliday, M.A.K. & Hasan, R. (1976). *Cohesion in English*. London: Longman.
Hancke, J., Vajjala, S., Meurers, D. (2012). Readability classification for German using lexical, syntactic, and morphological features. *Proceedings of COLING 2012*. 1063-1080.
Heilman, M., Collins-Thompson, K., Callan, J., & Eskenazi, M. (2007). Combining Lexical and Grammatical Features to Improve Readability Measures for First and Second Language Texts. In *Proceedings of HLT-NAACL'07*. 460–467.
Heilman, M., Collins-Thompson, K., & Eskenazi, M. (2008). An Analysis of Statistical Models and Features for Reading Difficulty Prediction. In *ACL 2008 BEA Workshop on Innovative Use of NLP for Building Educational Applications*.
Heilman, M., Collins-Thompson, K., Eskenazi, M., Juffs, A., & Wilson, L. (2010). Personalization of reading passages improves vocabulary acquisition. *International Journal of Artificial Intelligence in Education*, 20(1), 2010.
Honkela, T., Izzatdust, Z., & Lagus, K. (2012). Text mining for wellbeing: Selecting stories using semantic and pragmatic features. In *Artificial Neural Networks and Machine Learning–ICANN 2012*. Springer Berlin Heidelberg. 467-474.
Fernández Huerta, J. (1959). Medidas sencillas de lecturabilidad. *Consigna* (214): 29-32.
Jameel, S., Lam, W., & Qian, X. (2012). Ranking Text Documents on Conceptual Difficulty using Term Embedding and Sequential Discourse Cohesion. *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*. 145-152.
Kandel, L. & Moles, A. (1958). Application de l'Indice de Flesch à la langue français. *Cahiers d'Etudes de Radio-Television*, 19, 253-274.
Kanungo, T. & Orr, D. (2009). Predicting the readability of short web summaries. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining (WSDM '09).* ACM, New York, NY, USA, 202–211.
Kate, R. J., Luo, X., Patwardhan, S., Franz, M., Florian, R., Mooney, R. J., Roukos, S. & Welty, C. (2010). Learning to Predict Readability using Diverse Linguistic Features. *23rd International Conference on Computational Linguistics (COLING 2010).*
Kidwell, P., Lebanon, G., & Collins-Thompson, K. (2009). Statistical Estimation of Word Acquisition with Application to Readability Prediction. In *Proceedings of EMNLP'*09. 900–909.
Kidwell, P., Lebanon, G., & Collins-Thompson, K. (2011). Statistical Estimation of Word Acquisition with Application to Readability Prediction. *Journal of the American Statistical Association*. 106(493):21-30, 2011.
Kim, J.Y., Collins-Thompson, K., Bennett, P.N. & 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 (WSDM '12).* ACM, New York, NY, USA, 213–222.
Kincaid, J.P., Fishburne, R.P., Rogers, R.L., & Chissom, B.S. (1975). Derivation of New Readability Formulas (Automated Readability Index, Fog Count, and Flesch Reading Ease formula) for Navy Enlisted Personnel. Research Branch Report 8-75. Chief of Naval Technical Training: Naval Air Station Memphis.
Kireyev, K., & Landauer, T.K. (2011). Word Maturity: Computational Modeling of Word Knowledge. *Proceedings of ACL 2011*, 299-308.

Kittur, A., Chi, E.H., & Suh, B. (2008). Crowdsourcing User Studies with Mechanical Turk. *Proceedings of the 26th Annual ACM Conference on Human Factors in Computing Systems (CHI '08). ACM, 453-456.*

Klare, G.R. (1963). *The Measurement of Readability*. Ames, IA. Iowa State University Press.

Landauer, T. K., Kireyev, K., & Panaccione, C. (2011). Word Maturity: A New Metric for Word Knowledge. *Scientific Studies of Reading*, 15(1), 92-108.

Landauer, T. K., Kireyev, K., & Panaccione, C. (2009). A New Yardstick and Tool for Personalized Vocabulary Building. *BEA Workshop on Innovative Use of NLP for Building Educational Applications*. http://www.cs.rochester.edu/~tetreaul/naacl-bea4.html#program

Lau, T. P. (2006). Chinese Readability Analysis and Its Applications on the Internet. CUHK, Masters Thesis, Hong Kong, 2006.

Lennon, C. & Burdick, H. (2004). The Lexile Framework as an Approach for Reading Measurement and Success. Technical Report. Metametrics, Inc. April 2004. http://www.lexile.com/research/1/ (Retrieved Dec. 10, 2013)

Malvern, D. & Richards, B. (2012). Measures of Lexical Richness. *Encyclopedia of Applied Linguistics*, Blackwell Publishing Ltd.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society, Series B.* Vol. 42, No. 2, 109-142.

Mitchell, J.V. (1985). The Ninth Mental Measurements Yearbook. Lincoln, Nebraska: Univ. of Nebraska Press.

Nandhini, K. & Balasundaram, S.R. (2011). Improving readability of dyslexic learners through document summarization. In *Technology for Education (T4E), 2011 IEEE International Conference on. IEEE*, 246–249.

Nelson, J., Perfetti, C., Liben, D., Liben, M. (2012). Measures of Text Difficulty: Testing their Predictive Value for Grade Levels and Student Performance. Technical Report submitted to the Gates Foundation. Feb. 1, 2012. URL: http://achievethecore.org/content/upload/nelson_perfetti_liben_measures_of_text_difficulty_research_ela.pdf

Paivio, A., Yuille, J.C. & Madigan, S.A. (1968). Concreteness, Imagery, and Meaningfulness: Values for 925 Nouns. *Journal of Experimental Psychology*, 76, 1, Part 2 (1968), 1–25.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2 (1-2), 1-135.

Pilán, I., Volodina E., & Johansson, R. (2014). Rule-based and machine learning approaches for second language sentence-level readability. *BEA Workshop* 2014.

Pitler, E. & Nenkova, A. (2008). Revisiting readability: a unified framework for predicting text quality. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '08)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 186–195. http://dl.acm.org/citation.cfm?id=1613715.1613742

Rello, L., Saggion, H., Baeza-Yates, R., Graells, E. (2012). Graphical schemes may improve readability but not understandability for people with dyslexia. *Proceedings of NAACL-HLT 2012*.

Richardson, J.T. & E. (1975). Imagery, concreteness, and lexical complexity 2, Vol. 27. Psychology Press, 211–223.

Russell, D.M. (2011). SearchReSearch: Search by reading level [Web log post]. Retrieved from http://searchresearch1.blogspot.com/2011/02/search-by-reading-level.html

Schwarm, S.E. & Ostendorf, M. (2005). Reading level assessment using support vector machines and statistical language models. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL '05)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 523–530.

Sato, S., Matsuyoshi, S., & Kondoh, Y. (2008). Automatic Assessment of Japanese Text Readability Based on a Textbook Corpus. *Proceedings of LREC'08*, 2008.

Si, L. & Callan, J.P. (2001). A Statistical Model for Scientific Readability. In *Proceedings of CIKM'01*. 574–576.

Sitbon L., & Bellot, P. (2008). A readability measure for an information retrieval process adapted to dyslexics. In *Second international workshop on Adaptive Information Retrieval (AIR 2008)*. 52–57.

Sjöholm, J. (2012). Probability as readability: A new machine learning approach to readability assessment for written Swedish; Masters Thesis, Linköpings universitet, 2012.

Sung, Y. T., Chen, J. L., Cha, J. H., Tseng, H. C., Chang, T. H., & Chang, K. E. (2014). Constructing and validating readability models: the method of integrating multilevel linguistic features with machine learning. *Behavior Research Methods*, 1-15.

Stenner, A. J., Burdick, H., Sanford, E. E. & Burdick, D. S. (2007). The Lexile Framework for Reading Technical Report. MetaMetrics, Inc.

Tan, C., Gabrilovich, E.., & Pang, B. (2012). To Each His Own: Personalized Content Selection based on Text Comprehensibility. In *Proceedings of the 5th ACM International Conference on Web Search and Data Mining*, February 2012.

Tanaka, S., Jatowt, A., Kato, M.P., & Tanaka, K. (2013). Estimating content concreteness for finding comprehensible documents. In *Proceedings of WSDM'13*. 475–484.

Tanaka-Ishii, K., Tezuka, S., & Terada, H. (2010). Sorting by readability. *Computational Linguistics*, 36(2) 203-227.

Todirascu, A., François, T., Gala, N., Fairon, C., Ligozat, A. L., & Bernhard, D. (2013). Coherence and Cohesion for the Assessment of Text Readability. *Natural Language Processing and Cognitive Science*, 11.

Vapnik, V.N. (1995). The nature of statistical learning theory. Springer-Verlag New York, Inc.

Vajjala, S., & Meurers, D. (2012). On improving the accuracy of readability classification using insights from second language acquisition. *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*. ACL. 163-173.

Vajjala, S., & Meurers, D. (2014). Readability Assessment for Text Simplification: From Analyzing Documents to Identifying Sentential Simplifications. *ITL International Journal of Applied Linguistics*, Sept. 2014.

von Ahn, L. & Dabbish, L. (2008). Designing games with a purpose. *Communications of the ACM* 51, 8 (August 2008), 58-67. DOI=10.1145/1378704.1378719

Vor Der Brück, T. & Hartrumpf, S. (2007). A Semantically Oriented Readability Checker for German. *Proceedings of the 3rd Language & Technology Conference*, 270–274. Poznan, Poland. October 2007.

Wang, Y. (2006). Automatic Recognition of Text Difficulty from Consumers Health Information", *IEEE Symposium on Computer-Based Medical Systems*, Los Alamitos, CA, USA, IEEE Computer Society, 131–136.

Wiemer-Hastings, K., Krug, J., & Xu, X. (2001). Imagery, Context Availability, Contextual Constraint, and Abstractness. In *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*. Erlbaum, 1134–1139.

Zakaluk, B.L., & Samuels, S.J. (1988). Readability: its past, present and future. International Reading Association.

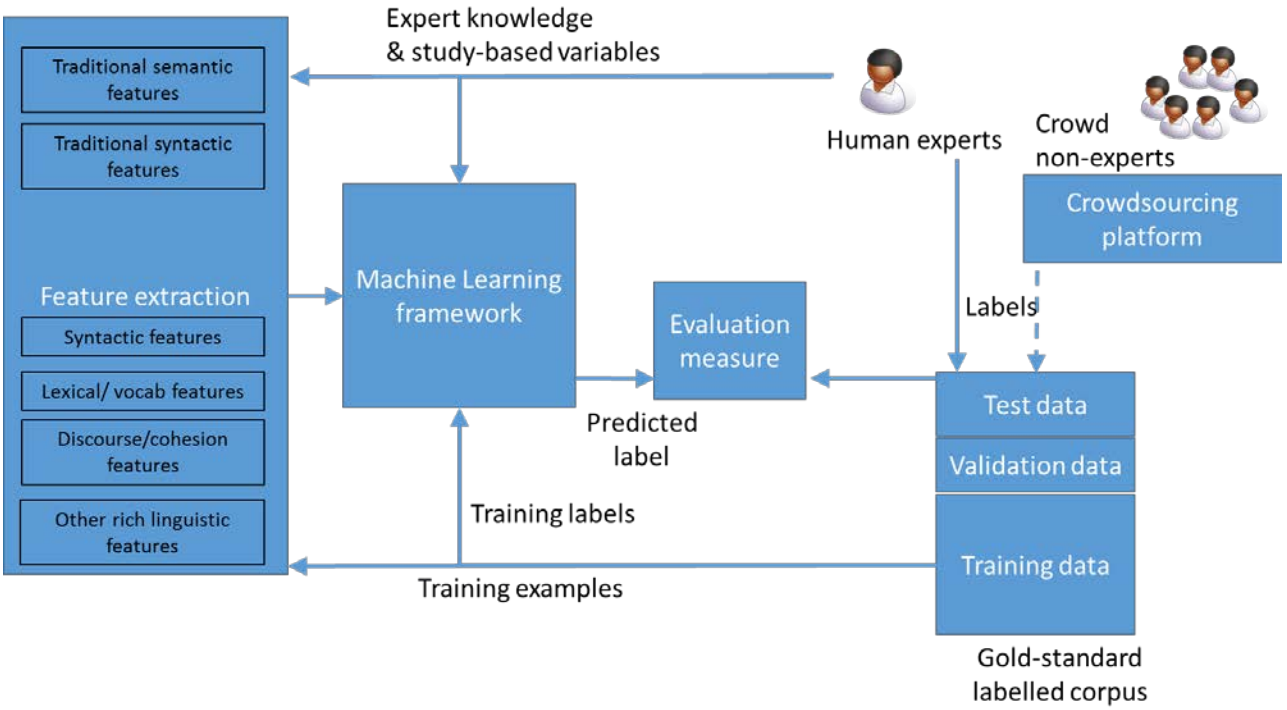


Figure 1: Overview of a typical computational reading difficulty estimation pipeline.

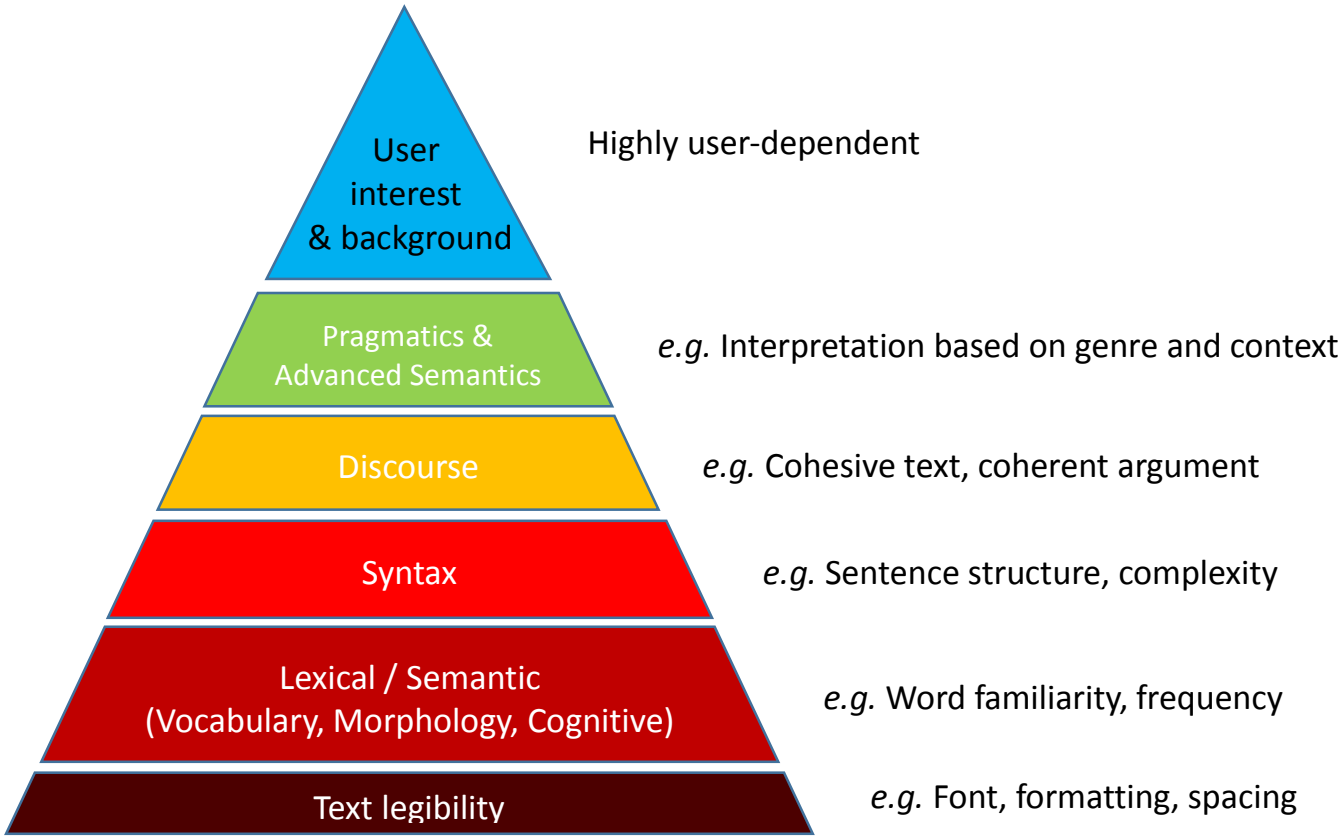


Figure 2: Key aspects of text readability, ordered from lowest level (text legibility) to highest level (user interest and background). These levels are one way to categorize the types of features used by text readability measures for automated assessment.

Lexical/semantic difficulty:

- Average number of syllables per word
- Out-of-vocabulary rate relative to a large corpus
- Type-token ratio: the ratio of unique terms to total terms observed
- Ratio of function words (compared to a general corpus in the target language)
- Ratio of pronouns (compared to a general corpus in the target language)
- Language model perplexity (comparing the text to generic or genre-specific models)

Syntactic difficulty:

- Average sentence length (in words or tokens)
- Proportion of incomplete parses
- Parse structure features:
    - Average parse tree height
    - Average number of noun phrases per sentence
    - Average number of verb phrases per sentence
    - Average number of subordinate clauses per sentence

Figure 3: Examples of typical lexical and syntactic features used for reading difficulty prediction, from Schwarm and Ostendorf (2005) and Kate et al. (2010).

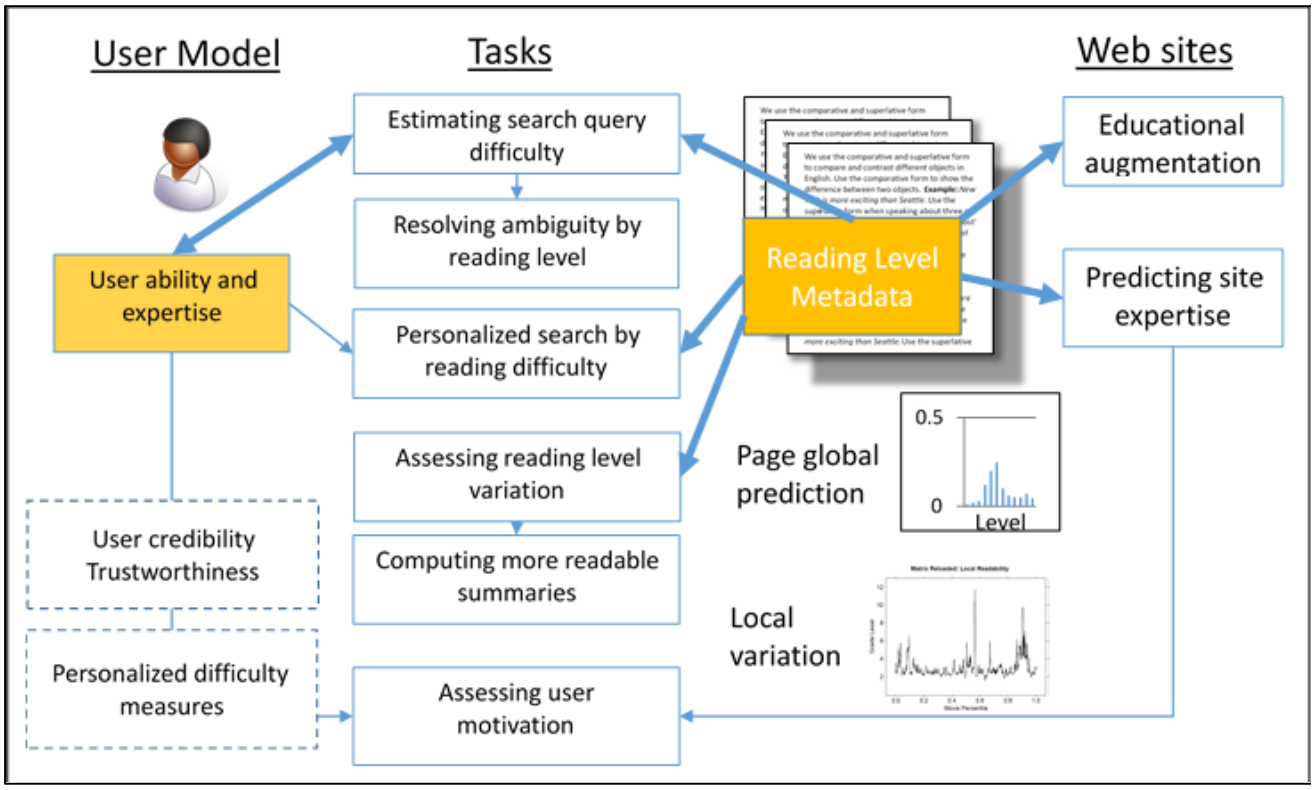


Figure 4: How computing metadata with readability estimates for Web pages enables a surprisingly wide variety of tasks and applications.

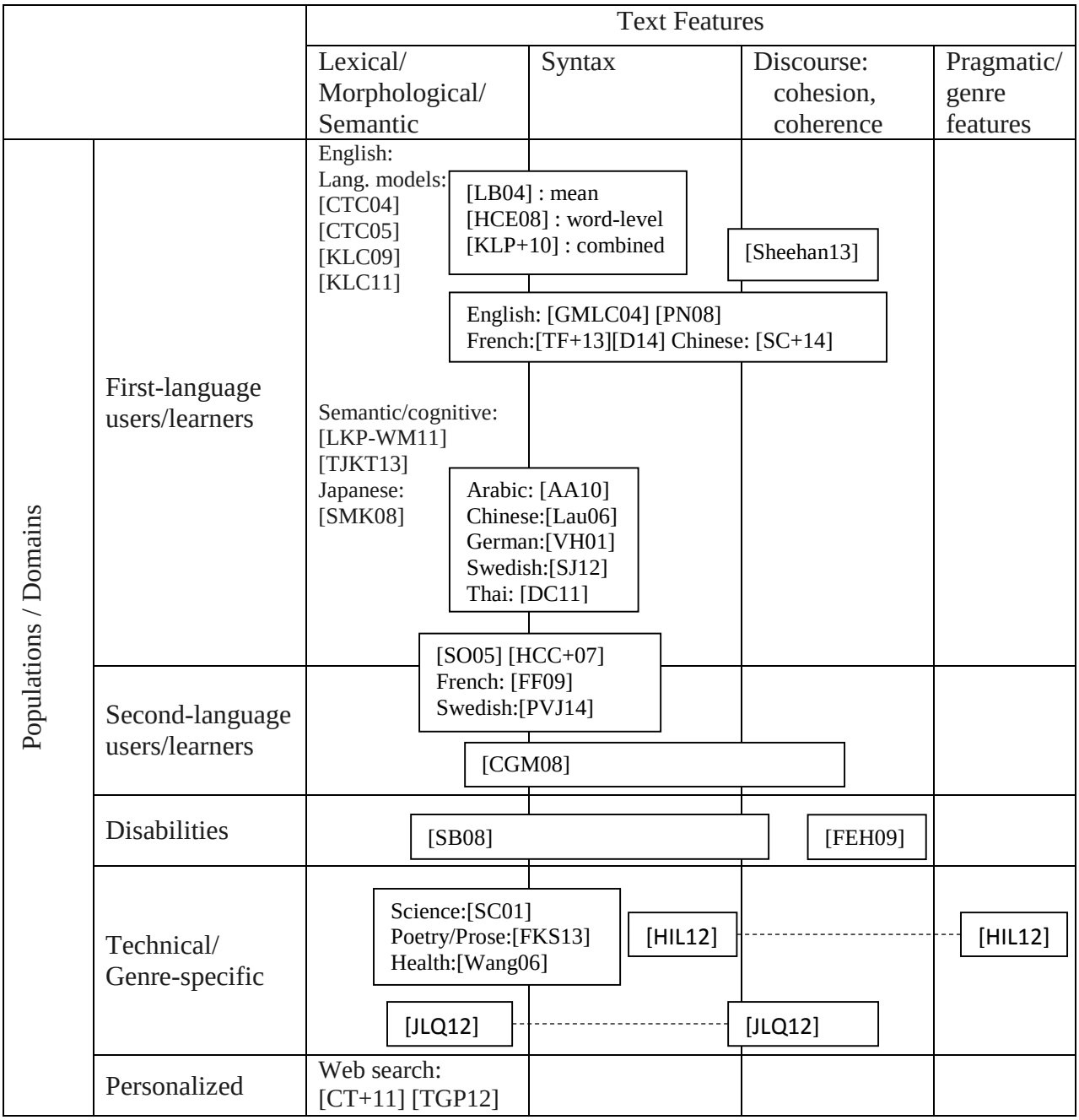


Figure 5: Visual summary of representative literature covered in this article that has introduced new automated readability assessment methods for different target populations. Papers (shown by citation key) have been classified in the horizontal direction according to the primary type or combination of features used to predict readability, and in the vertical direction by primary population or task target. The citation key concatenates the first letters of up to three initial authors last names (upper case) and appends the year, e.g. [FEH09] represents the 2009 paper of Feng, Elhadad, and Huenerfauth. ('+' means et al.) In case of ambiguity, additional lower-case letters are added from the first author's name.