Computational Assessment of Text Readability: A Survey of Past, Present, 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: July 2, 2014

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, machine learning.

Computational Assessment of Text Readability: A Survey of Past, Present, and Future Research

1. Introduction

Text readability is 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 be syntactic, semantic, graphical, discourse-based, or follow 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 matching people's information needs to the right type of content, 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. This 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. In this survey, we focus less on the visual aspects of readability, such as font size, which are objectively measurable, and more on subjective aspects of comprehension difficulty of the text. Thus, we sometimes use the phrases *text difficulty* or *reading difficulty* synonymously with 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. Most notable, however, is the key role that readability assessment continues to play in specific fields where ease of accessibility to critical information is especially important. These include finding educational material of the right difficulty for students in textbooks, on the Web, and other sources; calibrating public and private health information so that it is broadly understandable by the general public and individual patients, in the form of medical instructions, questionnaires, pamphlets and the like; producing effective user manuals and other documentation; creating informative and easy-to-understand Web sites and forms for critical government services; and assisting the world's information needs via the World Wide Web, social media and associated 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 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. 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 by presenting from a computer science and computational linguistics perspective. Recent education- or psychology-oriented reviews of readability measures, e.g. by Benjamin (2012) provide some related material, but the present work looks at a broader set of applications of which education is but one example. This means we focus less on issues such as classroom validation and making recommendations for school use – although these are by no means less critical - and more on how these methods actually work along with trends in likely advances in core text representations and algorithms. This article, being more recent, describes important additional methods such as Word Maturity (Landauer et al, 2012) not previously covered. Finally, based on my own literature survey as well as expensive experience developing both 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 automated 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 Chall (1958) and Klare (1963). *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 it has, and a widely-used proxy variable for a sentence's syntactic difficulty, its length in words. Second, the ordering of words and sentences is treated as exchangeable: the semantic variables are averaged over

all words, and syntactic variables averaged over all sentences, regardless of ordering. Thus, 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 makes sense for many traditional texts: vocabulary, and thus lexical features, are known to account for at least 80% of the variability in readability scores for traditional texts (Chall, 1958), while syntactic features account for much of the rest. 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 hundreds of similar variants that have been developed. (A study by Mitchell (1985) reviewed 97 different reading comprehension tests, although few of these have gained wide use.) 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 estimate the semantic difficulty component using a reference word list or standard corpus: these are termed 'vocabulary-based' traditional readability measures. A word is modeled as 'difficult' or 'unfamiliar' if it does not occur in the reference word list, or occurs with low frequency in the standard corpus. This word unfamiliarity variable then forms the semantic component of the traditional measure instead of more surface measures such as syllable count. Widely-used vocabulary-based traditional measures include the Lexile measure (Lennon & Burdick, 2004) the Revised Dale-Chall formula (Chall and Dale, 1995) and the Fry Short Passage measure (Fry, 1990). Lexile (version 1.0) uses the Carroll-Davies-Richman corpus (Carroll et al., 1971); Dale-Chall uses the Dale 3000 word list of words familiar to 80% of U.S. fourth-graders; and the Short Passage Measure uses Dale & O'Rourke's Living Word Vocabulary of 43,000 types (Dale and O'Rourke, 1981). All of these 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 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, their reliance on a small number of summary text features is both a strength and a weakness: simple formulas are generally easier to implement, but the same simple formulas have a basic inability to model the semantics of vocabulary usage in context.

Finally, and most importantly, 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 as the reader's prior knowledge and language skills 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 the opportunity to exploit the growing resources of content and difficulty data, 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 both of 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. Francois [FF09] has called this the 'AI' approach to readability. These new approaches typically combine a rich representation of the text being evaluated, using a variety of linguistic features, with more sophisticated prediction models based on machine learning. Like traditional readability formulas based on linear regression, 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 variables that characterize documents at different reading levels. These models in turn 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. 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, e.g. 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 to be extracted or 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, with these feature definitions and the gold standard corpus, a machine learning model learns how to predict the correct label for a text, given the set of extracted feature values. The model is learned by using randomly-selected subsets of the gold standard corpus as training data (typically 70%), to validate the choice of optimal parameters (20%) and as a final test set (10%). Features are extracted from the training examples, and the model is shown these feature vectors along with their corresponding gold standard label. The model parameters are adjusted using the validation set to minimize some measure of prediction error against the gold standard labels, such as Root Mean Squared Error (RMSE). Finally, the model is applied to the previously unseen test set to estimate how well the prediction model is likely 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 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 *why* a text is considered difficult or a good match for a student.

Because many factors can influence comprehension, assigning a 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, particularly 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). In a more recent 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:

- Lexical: rare, unfamiliar or ambiguous words.
- Morphological: rare or more complex morphological particles.
- Syntax: grammatical structure.
- Semantics: use of unusual senses, idioms, or subtle connotation.
- Discourse: inter-sentence relationships; explicit, clear argument structure of text.
- Cohesion: use of discourse connectives to clarify relationships or transitions.
- Pragmatic: language influenced by genre, e.g. sarcasm.
- Conceptual: domain or world knowledge required to comprehend a text.

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 vs. feature selection in readability prediction effectiveness.

3.2 Text features for readability assessment

Lexical features. Reflecting the importance of vocabulary in readability, *lexical features* capture attributes associated with the difficulty or unfamiliarity of 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. A particular readability prediction model could either use thousands of individual lexical feature values as input, 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 lexical diversity statistic known as the *type-token ratio*. The idea is that more advanced texts are authored using a larger vocabulary and will 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 of Collins-Thompson and Callan (2004) generalized this vocabulary-based approach, in which multiple language models are built automatically from training data – typically one for each grade level to be predicted – that capture more fine-grained information about vocabulary usage for each word. 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 statistical language models, 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. A key additional ability of Word Maturity, however, 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 at early stages, but acquires additional senses and subtleties of meaning, such as "surveillance device" in later stages. 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 simple forms of personalized readability measures.

The Word Maturity measure uses Latent Semantic Analysis to model the richness of contexts in which a word appears over time. An intermediate corpus is created for each learning stage to be modeled (e.g. grades). Next, the LSA space of that corpus is computed. A word at a particular grade or learning stage is represented by its LSA vector, which roughly correspond to the spectrum of topics in which it is used. A word's feature vector is also computed for a full, adult-level 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 in the 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 a beta version (as of this writing) of what they term the Reading Maturity Metric (RMM) that includes Word Maturity as one element, (http://www.readingmaturity.com/rmm-web/) as part of a range of computational linguistic features to assess syntactic complexity, coherence, and structural features of the text.

Syntactic features. Syntactic complexity is known to be associated with longer processing times in comprehension (Gibson, 1998) and thus factors significantly into automated readability assessment. Current 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 automatic, context-free parsers of sentences. Figure 3 shows a list of typical syntactic features derived from shallow and deep parsing.

[Insert Figure 3 here]

These 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. Example of these types of advanced syntactic features

may be found in (Schwarm & Ostendorf, 2005), (Heilman et al. 2007) or (Kate et al, 2010).

Higher-level cognitive-based text features. A number of studies have shown that more cohesive texts are less difficult to read. Texts or Web pages containing well-structured and thematically arranged content, clear section names, and other content arrangement techniques should be on average more readable than texts without this kind of content presentation. Yet cohesiveness is not captured by traditional readability formulas like Flesch-Kincaid. Newer automated assessment measures have attempted to remedy this issue by adding higher-level cohesion-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 text readability. 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.

Advances in computational linguistics, starting in the late 1970s, have made it possible to extract a variety of important new language features from textual material. Coh-Metrix [GMLC04] 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 computes many of the semantic and syntactic features mentioned here, as well as a rich set of lexical features such as a word's age of acquisition, concreteness, imagability, and degree of polysemy. This rich feature set in turn has been used in numerous studies to create readability measures such as those from (Crossley et al. 2008) for second-language learners. The recent work of Vajjala and Meurers (2012) is another example of research that is attempting to use insights from psycholinguistics and language acquisition to drive the development and investigation of new, cognitively-motivated readability features.

In general, we believe the full potential of natural language processing to capture features of more subtle, deeper levels of difficulty still has much potential. Features that involve the highest levels of text understanding, such as pragmatics, subtle semantics, and world knowledge are still very much in the early stages of exploration.

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, readability prediction can be treated as a form of *classification* task (with ordered or unordered category levels) or *regression problem* (with continuous-valued levels). In either case, the output variable is the readability level and the input variables are the set of feature values computed from the text as described above. Some studies, e.g. (Pitler & Nenkova, 2009) have also treated text readability as a *ranking* problem: that is, instead of predicting levels for individual documents, they predict the relative difficulty of pairs of documents. This is a natural and useful approach in cases where relative ordering is acceptable and an absolute prediction is not needed.

Heilman et al [HCE08] investigated how the choice of measurement scale affected prediction accuracy, and found that the most effective predictions of reading difficulty (measure in terms of correlation, RMSE, and accuracy) resulted from using a proportional-odds 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 tried learning frameworks such as Gaussian process regression, decision trees and support vector regression.

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. (Francois 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 best prediction performance used both classic and non-classic features. Depending on the evaluation measure used, support vector machines [V95] outperformed linear regression in accuracy, but had comparable explanatory power in terms of outcome variability.

A general conclusion that we can draw after reviewing dozens of studies using different combinations of learning frameworks and feature sets is that for readability prediction, the quality and selection of the features used as input to the learning framework usually matters more than the choice of the 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 model 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. Varying the learning framework used for prediction, they reported results using Gaussian Process Regression (0.7562), Decision Trees (0.7260), Support Vector Regression (0.7915), Linear Regression (0.7984), and Bagged Decision Trees (0.8173). While the model choice does affect performance, the variation in performance due to changing the learning model was smaller than the variation in changing the features. In our experience, this is typical of many machine learning studies for readability prediction.

Thus, other considerations 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 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 somewhat 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 company 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. One caveat: 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 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.) The 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 vs children, originally collected by (Barzilay and Elhaded, 2003). Similar two-level easy/difficult corpora are available for the Wikipedia (simplified English vs default English). (Pitler & Nenkova, 2008) annotated 30 Wall Street Journal articles for readability scores (in the range 1-5). A few domain-specific corpora are available, such as the corpus on math readability that contains 120 documents labeled on a difficulty scale from 1to 7 (available 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 significant, freely available, high-quality corpora for automated readability evaluation.

Evaluation measures. The rank order correlation (Spearman's *rho*) between the difficulty levels predicted by the readability measure for the reference texts, and the 'gold standard' difficulty levels (or other independent measure) provided for the same reference texts is one widely-used evaluation measure. 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: normalization of the readability scores that may be output from the measure (and in comparison across multiple measures), which may be on a very different scale compared to the gold-standard label, or with other measures, is not required. Thus, Spearman correlation is robust to outliers, and does 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. In all cases involving machine learning models trained from data, the technique of cross-validation (typically using 10 folds) is used to assess the likely variability and generalization error, by training on different randomly selected subsets of the training data, measuring the prediction error over the remaining test data, and averaging the prediction error over all cross-validation folds.

Evaluation results. How large are rank correlations for the more advanced commercial 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. 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 page recommendation, but also some surprising new capabilities like estimating user motivation during Web search, as I 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 such as (Crossley et al. 2008), who used a rich feature set that included syntactic sentence similarity, lexical co-referentiality, and word frequency, as computed by the Coh-Metrix computational tool. Schwarm and Ostendorf's work (2005) on general readability prediction was partly motivated by the need for tools in bilingual education. Their work used corpora that included simplified English texts but did not explicitly report results for L2 readers.

International Language Support

In general, readability assessment research has traditionally focused on studies in English first, with other languages adapting and extending those results: after the Flesch formula for readability of English text [F48] was published in 1948, a series of adaptations for European and other languages followed. For example, in 1958 Kandel and Moles published an adaptation of the Flesch Reading Ease formula (cf. [FF09]) for calculating the readability of French text, and in 1959 José Fernández Huerta published a corresponding formula for Spanish text that is still widely used [FH58]. A similar trend, where English-language development is followed by European and Asian languages, has been evident in the most recent machine learning approaches to readability. However, there are also notable differences compared to the development of traditional measures: adaption of automated methods beyond English has happened on a much more compressed time scale than traditional methods, and in particular, Asian languages have been much earlier adapters of improved computational methods.

Varying degrees of effort are needed to re-purpose machine learning-based automated assessment methods originally developed for 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 syntactic difficulty features derived from parse trees, requires NLP tools like 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 intriguing 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).

Languages as diverse as German [VH07], French [FF09], Arabic [AA10], Thai [DC11] and Swedish [SJ12] are among those recently applying new semantic resources and learning-based computational methods. 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/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 is critical: shorter words help comprehension. Finally, (Feng et al, 2009) developed and evaluated automated 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.

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.

In one example, the REAP system at Carnegie Mellon University (http://reap.cs.cmu.edu) 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 most effectively. In a controlled study, for example, using the underlying REAP search engine to personalize the topics of example material led to consistent gains in student performance in vocabulary acquisition [HCE+10].

In related work, Beinborn et al. [BZG12] 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 [DDT+13], 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

As we described earlier, traditional readability measures do not work well on Web content. One reason is the highly varied, non-traditional nature of content, from blog comments to captions from search engine result pages, to online advertising. Web pages also can contain images, video, 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 is especially true for Web search systems – a primary portal to people's access to information on the Web. Traditionally, search engines have ignored the reading difficulty of documents and reading proficiency of users in modeling relevance, ranking documents, and many other aspects of retrieval. This is evident with the increased interest in more effective search systems for children and students (Collins-Thompson et al, 2011). While addressing children's search needs 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.

Similarly, experts may not want tutorials and introductory texts and instead prefer material that is actually highly technical. Non-native language speakers also form a significant population of users who could benefit from improvements in information retrieval that account for reading level.

Enriching Web pages with readability metadata has led to a variety of new and sometimes surprising applications. 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, site, and author expertise (Kim et al. WSDM 2012).

[Insert Figure 4 here]

Also unlike traditional texts, Web pages have valuable additional sources of information, such as the set of links to and from the page, 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. For example, (Gyllstrom and Moens 2010) proposed AgeRank, an algorithm that provides a binary labeling of Web documents according to its appropriateness for children vs. 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 as it is known, or the nature of user interactions with content relative to reading difficulty. 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 level. Some recent work has begun to study Web readability via user interactions via query logs: Torres et al. [12] 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' Open Directory Project top-level category. 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. One approach is to have users self-identify their level of proficiency. This is the approach Google has used in their recent deployment of an Advanced Search feature to filter results by Low, Medium, and High levels of difficulty. 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. Early work on automatically estimating a reading proficiency profile for a specific user from their interaction with the Web is by (Collins-Thompson et al. 2011) and then by (Tan et al, 2012). We expect that existing learning algorithms could be applied based on such observations as the reading level of past (satisfied) clicked documents; semantic or syntactic features of current and past queries; or 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 also forsee the need for models that capture expertise on specific topics, in addition to general reading proficiency.

Performing Web search by ranking results according to their readability level aims at reducing the 'gap' between the user's reading proficiency distribution and a document's reading level distribution. 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. Exploring reliable methods for modifying existing rankings based on a user reading proficiency profile is an area of current research.

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 retrieval 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 the 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, such as [HCC+07], [CGM08], and [JLQ12] 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, few non-English readability models have incorporated features of higher-level text structure - 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 English-language domains that include technical or scientific writing and poetry/prose. Third, there has been little published work on learning more fine-grained, individual models of reading expertise.

The closest relevant work so far published in that area has been related to Web search [CT+11][TGP12], where user search behavior, such as queries issued and documents clicked, has been used to estimate personalized models of user interest and expertise. 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 do not reflect an individual user's context or knowledge. Moreover, users are dynamic individuals (usually) and have expertise and interests that evolve over time, as well as 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**. Supervised machine learning models require data to learn from – this is both their strength and weakness. To obtain labeled data, the use of human computation and crowdsourcing are promising avenues just beginning to be explored. 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 our 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 easy 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 *first* before they can understand a second concept will prove invaluable sources of assistance for many tasks. Health informatics is one area in particular where research on capturing and exploiting deep domain knowledge will lead to much more effective readability metadata for helping users.

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 linguistics community that standardizes task and evaluation criteria, like those already organized annually for other computational linguistics 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.

I 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 [CTC04a]. In a related direction, Agrawal et al. use estimates of syntactic complexity and key concepts to identify difficult sections of textbooks that could benefit from better exposition [AGK+11] 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) included a 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 that are need of additional explanation or augmentation (e.g. [AGK+11]).

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. In one example, a recent study by (Cole et al., 2012) showed that a user's level of domain knowledge could be inferred 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 of readability 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 had resulted in a subsequent lack of labeled corpora, because 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, can 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 (as in 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 for that optimize for efficient crowdsourcing of readability labels or features are likely to be a fruitful tool, and area for on-going research in their own right.

7 Conclusion

Automatic readability assessment promises 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 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 difficulties in developing and evaluating 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 methods for 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

[ALK+11] J. Abedi, S. Leon, J. Kao, R. Bayley, N. Ewers, J. Herman, K. Mundhenk. 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

[AGK+11] Rakesh Agrawal, Sreenivas Gollapudi, Anitha Kannan, and Krishnaram Kenthapadi. 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.

[APJ+11] Kouichi Akamatsu, Nimit Pattanasri, Adam Jatowt, and Katsumi Tanaka. 2011. Measuring Comprehensibility of Web Pages Based on Link Analysis. In Web Intelligence.

[AA10] Al-Khalifa, H. S. and Al-Ajlan, A. A. 2010. Automatic Readability measurements of the Arabic text: An exploratory study. Arabian Journal for Science and Engineering, 2010, 35.

[BE03] Barzilay, R., Elhadad, N., (2003) Sentence Alignment for Monolingual Comparable Corpora, in Proceedings of EMNLP, 2003.

[Bat03] 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.

[Bec04] Shirley Ann Becker. 2004. A study of web usability for older adults seeking online health resources. ACM Transactions on Computer-Human Interaction (TOCHI) 11, 4 (2004), 387–406.

[BZG12] 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.

[Ben12] R. Benjamin (2012). Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. Educational Psychology Review, 24(1):63-88.

[CNPM13] 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.

[CBC+13] Xi Chen, Paul N. Bennett, Kevyn Collins-Thompson, and Eric Horvitz. 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.

[CGL+12] Cole, M.J., Gwizdka, J., Liu, C., Belkin, N.J., Zhang, X. Inferring user knowledge level from eye movement patterns. Information Processing and Management, 2012. DOI: http://dx.doi.org/10.1016/j.ipm.2012.08.004

[CT+11] Kevyn Collins-Thompson, Paul N. Bennett, Ryen W. White, Sebastian de la Chica, and David Sontag. 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.

[CTC04a] Kevyn Collins-Thompson and Jamie Callan. 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.

[CTC04b] Kevyn Collins-Thompson and James P. Callan. 2004c. A Language Modeling Approach to Predicting Reading Difficulty. In HLT-NAACL. 193–200.

[CTC05] Collins-Thompson, K. and Callan, J.; Predicting reading difficulty with statistical language models; Journal of the American Society for Information Science and Technology, 2005, 56, 1448-1462.

[CGM08] Crossley, S. A., Greenfield, J., & McNamara, D. S. (2008). Assessing text readability using cognitively based indices. TESOL Quarterly, 42 (3), 475-493.

[DC49] E. Dale and J.S. Chall. 1949. The concept of readability. Consciousness and Cognition 26(23) (1949).

[DC11] Daowadung, P. and Chen, Y.-H.; 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.

[DH+13] Orphée De Clercq, Veronique Hoste, Bart Desmet, Philip van Oosten, Martine De Cock, Lieve Macken. Using the Crowd for Readability Prediction. Natural Language Engineering. 1(1). Cambridge University Press, 2013.

[ESV11] Carsten Eickhoff, Pavel Serdyukov, and Arjen P de Vries. 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.

[ET87] Carol Sue Englert and Carol Chase Thomas. 1987. Sensitivity to text structure in reading and writing: A comparison between learning disabled and non-learning disabled students. Learning Disability Quarterly (1987), 93–105.
[FKL08] Timo Faaß, Lars Kaczmirek, and Alwine Lenzner. 2008. Psycholinguistic Determinants of Question Difficulty: A Web Experiment. In 7th International Conference on Social Science Methodology.
[FEH09] Lijun Feng, Noemie Elhadad, and Matt Huenerfauth. 2009. Cognitively motivated features for readability assessment. In The 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2009).
[FKS13] Flor, M.; Klebanov, B. B. and Sheehan, K. M.; Lexical Tightness and Text Complexity; Proceedings of the Second Workshop onNatural Language Processing for Improving Textual Accessibility, 2013.
[F09] François, T. L; 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, 20
[FF09] François, T. and Fairon, C. 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), Jeju, 466-477. 09, 19-27.
[FM12] François, T. and Miltsakaki, E.; 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.
[FM91] Ferguson, G., Maclean, J. (1991) Assessing the readability of medical journal articles: an analysis of teacher judgements. Edinburgh Working Papers in Linguistics. No. 2, pg 112-125. http://files.eric.ed.gov/fulltext/ED353790.pdf
[Gib98] Gibson, E. (1998) Linguistic complexity: locality of syntactic dependencies. *Cognition,* 68:1-76.
[GMLC04] A. C. Graesser, D. S. McNamara, M. M. Louwerse, and Z Cai. 2004. Coh-metrix: analysis of text on cohesion and language. Behavior Research Methods, Instruments, and Computers 36(2) (2004), 193–202.
[GM10] Karl Gyllstrom and Marie-Francine Moens. 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.
[HCC+07] Michael Heilman, Kevyn Collins-Thompson, Jamie Callan, and Maxine Eskenazi. 2007. Combining Lexical and Grammatical Features to Improve Readability Measures for First and Second Language Texts. In HLT-NAACL. 460–467.
[HCE08] Michael Heilman, Kevyn Collins-Thompson, and Maxine Eskenazi. 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.
[HCE+10] M. Heilman, K. Collins-Thompson, M. Eskenazi, A. Juffs, L. Wilson. "Personalization of reading passages improves vocabulary acquisition." International Journal of Artificial Intelligence in Education, 20(1), 2010.
[JLQ12] S. Jameel, W. Lam, X. Qian. 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. pg 145-152.

[KO09] Tapas Kanungo and David Orr. 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.

[KLP+10] Kate, R. J.; Luo, X.; Patwardhan, S.; Franz, M.; Florian, R.; Mooney, R. J.; Roukos, S. and Welty, C.; Learning to Predict Readability using Diverse Linguistic Features; 23rd International Conference on Computational Linguistics (COLING 2010), 2010.

[KLC09] Paul Kidwell, Guy Lebanon, and Kevyn Collins-Thompson. 2009. Statistical Estimation of Word Acquisition with Application to Readability Prediction. In EMNLP. 900–909.

[KLC11] P. Kidwell, G. Lebanon, K. Collins-Thompson. Statistical Estimation of Word Acquisition with Application to Readability Prediction. Journal of the American Statistical Association. 106(493):21-30, 2011.

[KCB+12] Jin Young Kim, Kevyn Collins-Thompson, Paul N. Bennett, and Susan T. Dumais. 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.

[KF+75] 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.

[KLWM11] Kirill Kireyev, Thomas K. Landauer: Word Maturity: Computational Modeling of Word Knowledge. ACL 2011: 299-308.

[KX01] Jan Krug and Xu. 2001. Imagery, Context Availability, Contextual Constraint, and Abstractness. In Proceedings of the 23rd Annual Conference of the Cognitive Science Society. Erlbaum, 1134–1139.

[LKP-WM11] Landauer, Thomas K.; Kireyev, Kirill; Panaccione, Charles. Scientific Studies of Reading, v15 n1 p92-108 2011.

[LKP-WM09] Thomas Landauer, Kirill Kireyev and Charles Panaccione. 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

[L06] Lau, T. P.; Chinese Readability Analysis and Its Applications on the Internet; CUHK, Masters Thesis, Hong Kong, 2006.

[LB04] Colleen Lennon and Hal Burdick. 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)

[M80] McCullagh, P. (1980) Regression models for ordinal data. *Journal of the Royal Statistical Society, Series B.* Vol. 42, No. 2, pp 109-142.

[NB11] K Nandhini and SR Balasundaram. 2011. Improving readability of dyslexic learners through document summarization. In Technology for Education (T4E), 2011 IEEE International Conference on. IEEE, 246–249.

[NP+12] 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

[PYM68] Allan Paivio, John C. Yuille, and Stephen A. Madigan. 1968. Concreteness, Imagery, and Meaningfulness: Values for 925 Nouns. Journal of Experimental Psychology 76, 1, Part 2 (1968), 1–25.

[PL08] Pang, B., Lee, L. 2008. Opinion mining and sentiment analysis. Foundations and trends in information retrieval 2 (1-2), 1-135.

[PN08] Emily Pitler and Ani Nenkova. 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

[RS+12] Luz Rello, Horacio Saggion, Ricardo Baeza-Yates, and Eduardo Graells. 2012. Graphical schemes may improve readability but not understandability for people with dyslexia. NAACL-HLT 2012 (2012), 25.

[RI75] John T.E. Richardson. 1975. Imagery, concreteness, and lexical complexity. 2, Vol. 27. Psychology Press, 211–223.

[R06] Randy Rumbo. 2006. English Composition 1: Using Specific and Concrete Diction. (2006). http://www2.ivcc.edu/rambo/eng1001/eng1001 diction.htm.

[SC91] P.J. Schwanenflugel. 1991. Why are Abstract Concepts Hard to Understand? 223–250.

[SO05] Sarah E. Schwarm and Mari Ostendorf. 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.

[SMK09] Sato, S.; Matsuyoshi, S. and Kondoh, Y.; Automatic Assessment of Japanese Text Readability Based on a Textbook Corpus. LREC'08, 2008.

[S13] Sheehan, K.M.; Measuring Cohesion: An Approach That Accounts for Differences in the Degree of Integration Challenge Presented by Different Types of Sentences; Educational Measurement: Issues and Practice, 2013.

[SC01] Luo Si and James P. Callan. 2001. A Statistical Model for Scientific Readability. In CIKM. 574–576.

[SB08] Laurianne Sitbon and Patrice Bellot. 2008. A readability measure for an information retrieval process adapted to dyslexics. In Second international workshop on Adaptive Information Retrieval (AIR 2008). Citeseer, 52–57.

[SJ12] Sjöholm, J.; Probability as readability:A new machine learning approach to readability assessment for written Swedish; Masters Thesis- Linköpings universitet, 2012.

[SYS+06] Rebecca L Sudore, Kristine Yaffe, Suzanne Satterfield, Tamara B Harris, Kala M Mehta, Eleanor M Simonsick, Anne B Newman, Caterina Rosano, Ronica Rooks, Susan M Rubin, and others. 2006. Limited literacy and mortality in the elderly: the health, aging, and body composition study. Journal of General Internal Medicine 21, 8 (2006), 806–812.

[SBS+07] Stenner, A. J., Burdick, H., Sanford, E. E. & Burdick, D. S. (2007). The Lexile Framework for Reading Technical Report. MetaMetrics, Inc.

[TGP12] Tan, C., Gabrilovich, E.., and Pang, B. 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.
[TJKT13] Shinya Tanaka, Adam Jatowt, Makoto P. Kato, and Katsumi Tanaka. 2013. Estimating content concreteness for finding comprehensible documents. In WSDM. 475–484.
[V95] Vladimir N. Vapnik. 1995. The nature of statistical learning theory. Springer-Verlag New York, Inc.
[VM12] 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.
[VH07] Tim Vor Der Brück and Sven Hartrumpf. A Semantically Oriented Readability Checker for German. Proceedings of the 3rd Language & Technology Conference, pp. 270–274. Poznan, Poland. October 2007.
[W06] Y. Wang, "Automatic Recognition of Text Difficulty from Consumers Health Information", IEEE Symposium on Computer-Based Medical Systems, Los Alamitos, CA, USA, IEEE Computer Society, 2006, pp. 131–136.
[W88] M D Wilson. 1988. MRC Psycholinguistic Database: Machine Readable Dictionary, Version 2. Behavioural Research Methods Instruments and Computers 20, 1 (1988), 6–11.

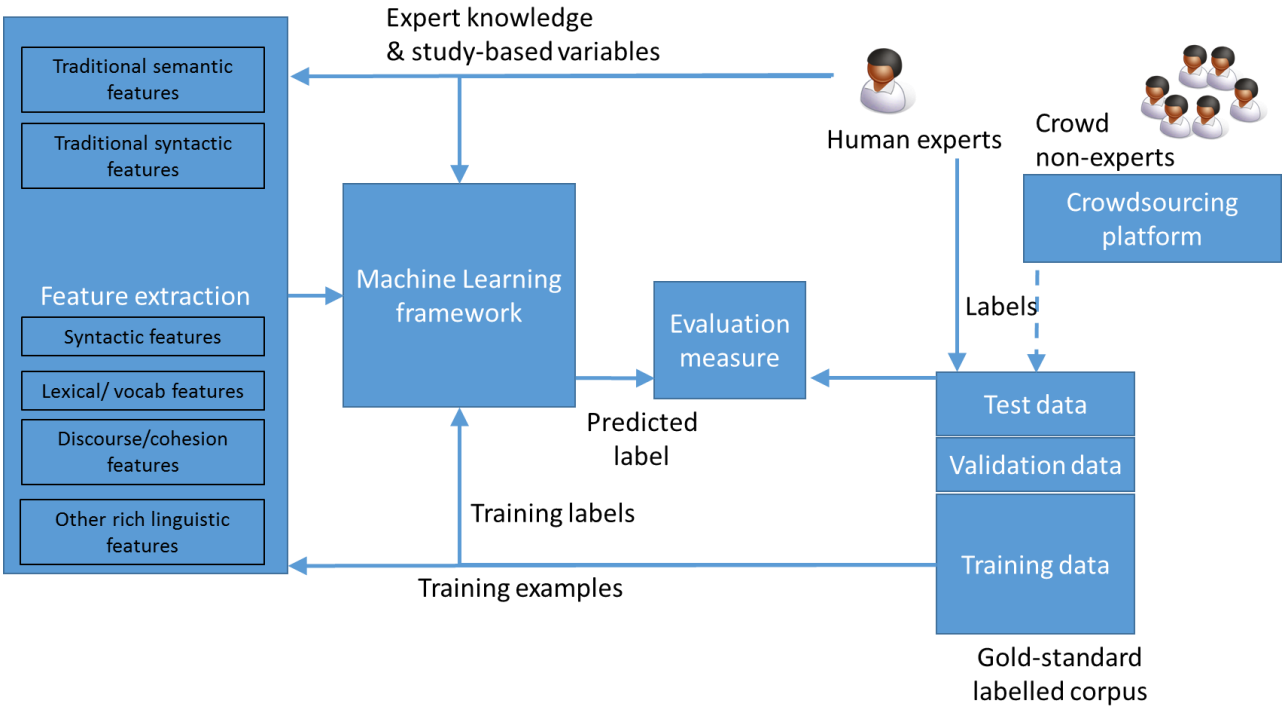

Figure 1: Overview of a typical computational reading difficulty estimation pipeline.

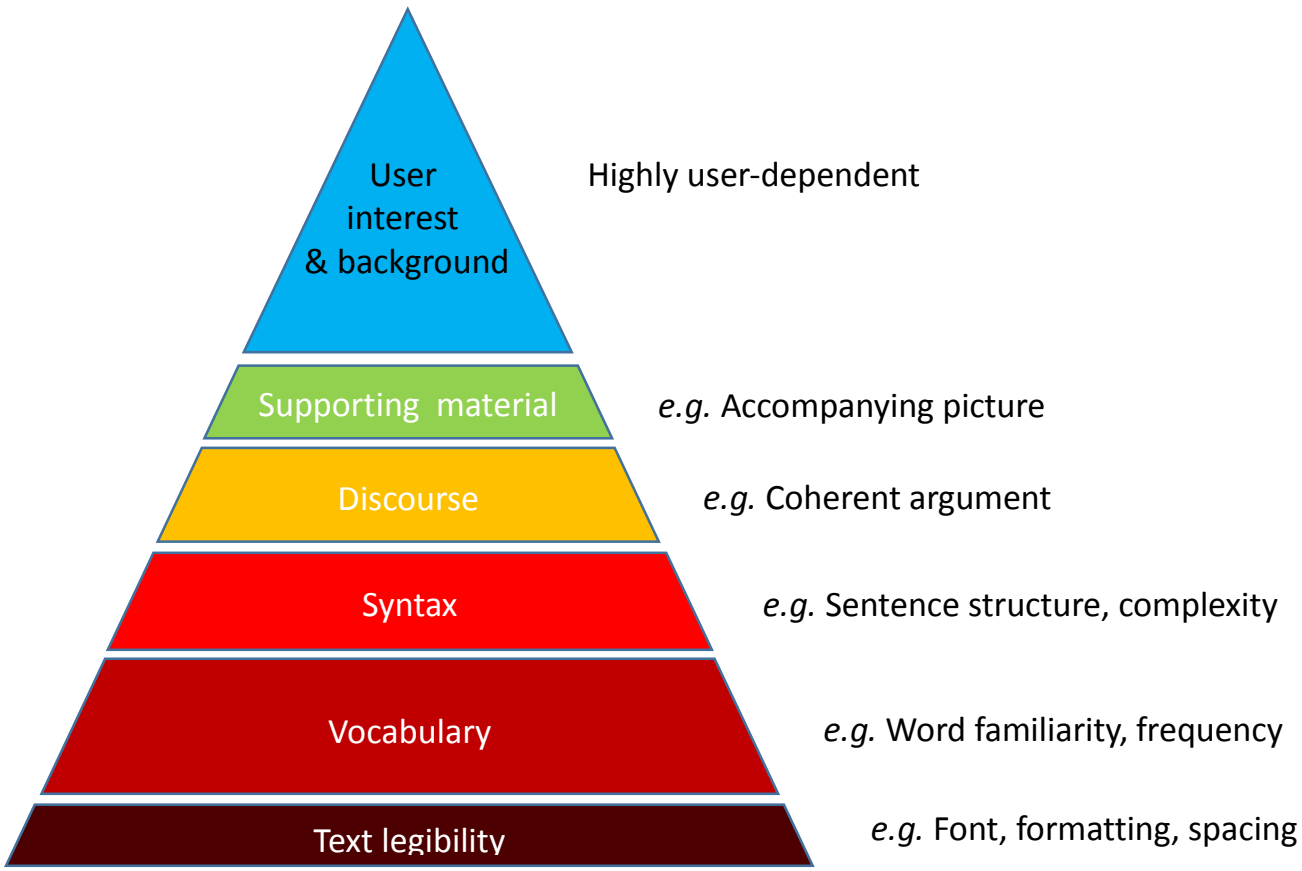

Figure 2: Key aspects of text readability, ordered from lowest level (text legibility) to highest level (user interest and background). These levels can serve to categorize the types of features used by text readability measures for automated assessment.

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 syntactic and lexical features used for reading difficulty prediction, from [SO05] and (Kate et al., 2010).

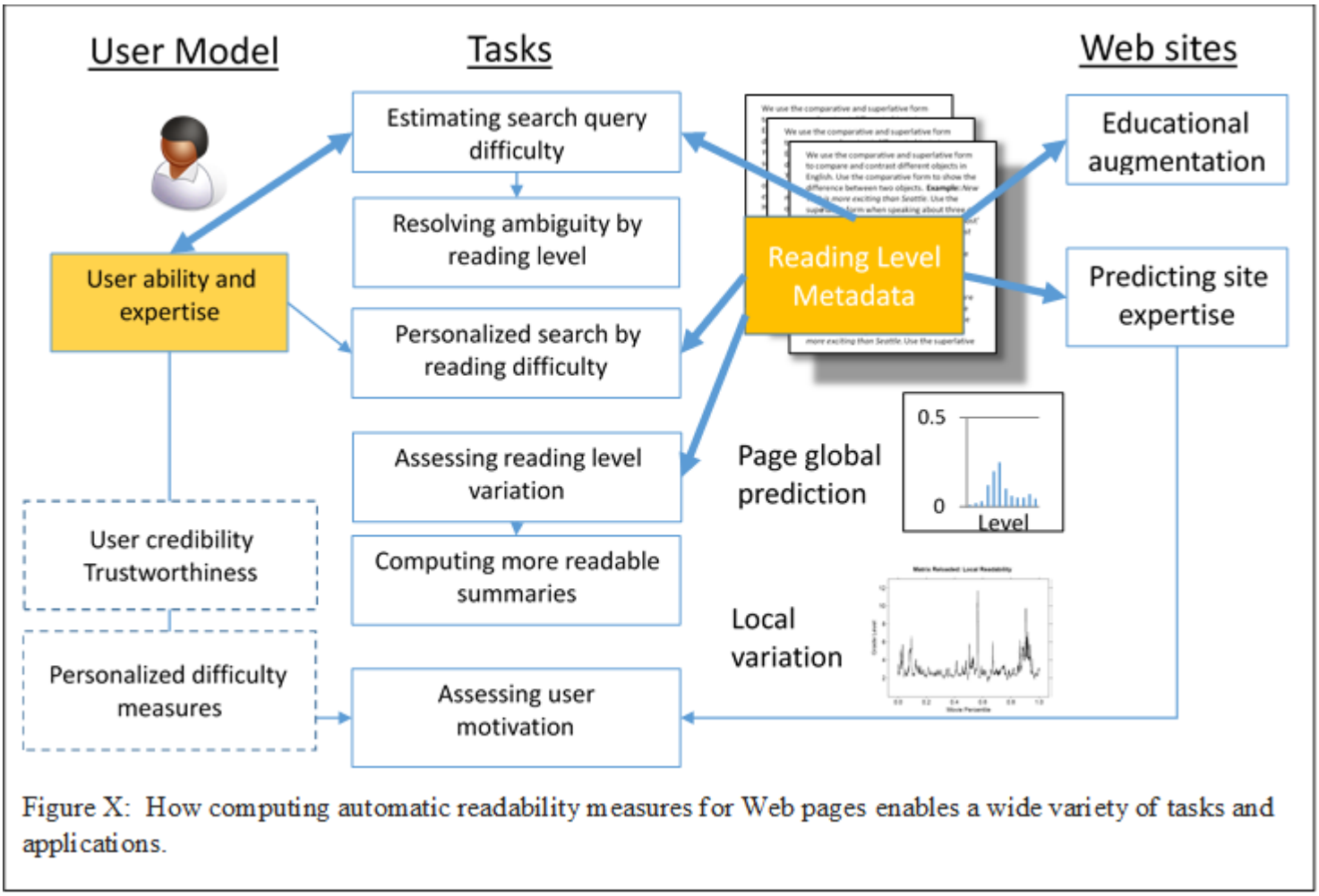

Figure X: How computing automatic readability measures for Web pages enables a wide variety of tasks and applications.

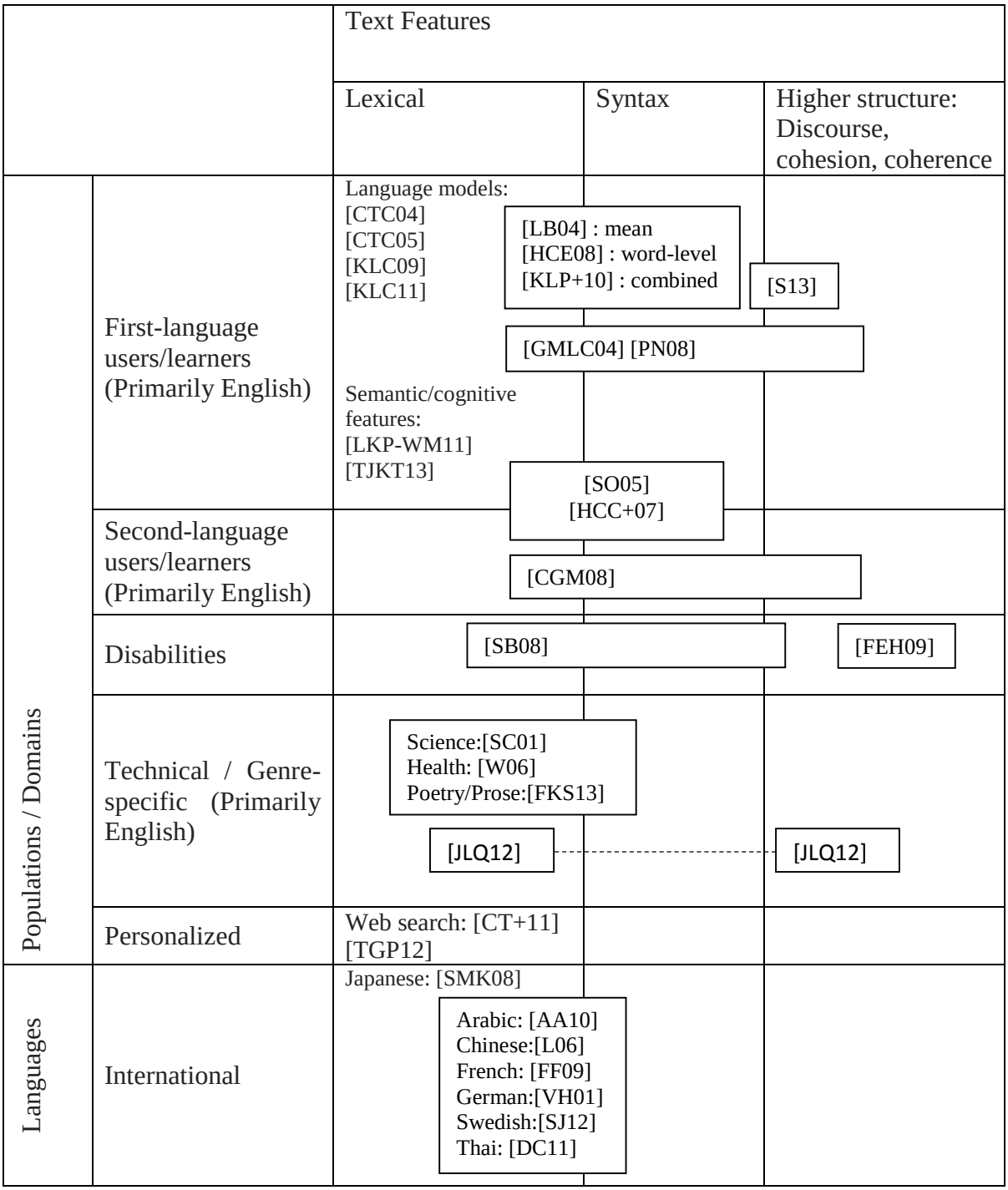

Figure 5: Visual summary of representative literature covered in this article that has introduced new automated readability assessment methods for different target populations. Papers (shown by citation key) have been classified in the horizontal direction according to the primary type or combination of features used to predict readability, and in the vertical direction by primary population or task target.