Towards fine-grained readability measures for self-directed language learning

Lisa Beinborn, Torsten Zesch and Iryna Gurevych

Ubiquitous Knowledge Processing Lab (UKP-TUDA)
Department of Computer Science, Technische Universität Darmstadt

Ubiquitous Knowledge Processing Lab (UKP-DIPF)
German Institute for International Educational Research

www.ukp.tu-darmstadt.de

Abstract

In this paper, we analyze existing readability measures regarding their applicability to self-directed language learning. We identify a set of dimensions for text complexity and focus on the lexical, syntactic, semantic, and discourse dimensions. We argue that for the purposes of self-directed language learning, the assessment according to the individual dimensions should be preferred over the overall readability prediction. Furthermore, due to the heterogeneity of the learners in such a setting, modeling the background knowledge of the learner becomes a critical step.

1 Introduction

Readability measures have a long history, especially in American education research (DuBay, 2004). The need for these measures is rooted in a very practical task: teachers search for texts that best fit the knowledge level of their students. According to Vygotsky's zone of proximal development (Vygotsky, 1978), the range of suitable texts that a learner can manage without help is very small. Texts that do not challenge the student easily lead to boredom, while overly complex language might lead to frustration when no tutoring is available. In order to prepare useful reading material for students, readability measures assign the most suitable school grade level to each text. Thus, the readability measures provide an approximation of the text complexity.

The existing readability measures have been developed for standard classroom teaching. As an alternative, *self-directed learning* has lately been on the rise. Self-directed learning refers to a learning setting that does not involve a teacher. It's emergence is closely related to the increased availability of educational material on the web. Students use online exercises for additional training, and companies have discovered digital courses as a flexible alternative to educate their employees. The main advantage of self-directed learning, as opposed to standard classroom education, is the focus on the independence and individuality of the learner (see table 1). The learner can work in her own rhythm independent of time slots or opening hours of institutions and can make her own decisions about the learning content and strategy. A readability measure for self-directed learning thus needs to account for the individual user profiles of learners.

A typical application for self-directed learning is *language learning* where exercises are usually coupled with a text that introduces new vocabulary or new grammatical constructions.[1] It is very important for the learning process that the text fits the proficiency level of the learner. This goal can be reached by applying readability measures which provide an objective analysis about the text complexity. It should be noted, though, that most readability measures have been developed for native speakers rather than for foreign language learners. However, the acquisition of the native language and the learning of a second language are very different processes (see table 2). A readability measure for language learning should take these differences into

[1] In this paper, we use the term language acquisition to refer to the process of acquiring the native language and language learning to the process of learning a foreign language.

Lisa Beinborn, Thorsten Zesch and Iryna Gurevych 2012. Towards fine-grained readability measures for self-directed language learning. *Proceedings of the SLTC 2012 workshop on NLP for CALL.* Linköping Electronic Conference Proceedings 80: 11–19.

	Classroom	Self-directed
Learner	Homogeneous group	Individual
Learning Mode	Teacher	Independent
Background Knowledge	Curriculum	Individual

Table 1: Differences between classroom learning and self-directed learning

	L1 acquisition	L2 learning
Learner	Young child	Unspecified
Learning Mode	Unstructured	Structured
Background Knowledge	No language knowledge	L1

Table 2: Differences between native language acquisition and second language learning

account.

Self-directed language learning gives the learner the opportunity to improve their command of a language on their own. Advanced learning systems can abstract from pre-defined curricula and adapt the content to the specific learner resulting in a very personalized learning setting (Karel and Klema, 2006). Such an advanced learning framework requires flexible technology that is able to react to the user feedback and continuously update the assumptions of the learner's knowledge.

To the best of our knowledge, the requirements for readability measures for self-directed language learning have not yet been studied in detail. Previous surveys give a historical overview of the evolution of readability measures in the classroom setting. DuBay (2004) introduces the most popular traditional approaches to readability in detail and also presents experimental readability studies. Benjamin (2011) evaluates readability measures according to their usability for teachers. Recently, progress in the field of text classification has led to a new perspective on readability measures not yet captured by previous surveys. Text features from various dimensions are taken into account and combined by supervised learning. In this paper, we present the different approaches to measuring readability and group the introduced text features according to their linguistic dimension. In addition, we discuss how the existing approaches can be adapted to other languages, to language learning, and to self-directed learning.

2 Dimensions of text complexity

A text can be difficult in several ways. A reader might for example know each word of the text, but still fail to capture the constructed meaning. In language learning, these differences are even more evident. If a text uses an unknown grammar construction, the learner will fail to comprehend the text regardless of the vocabulary used. In order to understand and predict the difficulties a learner might have with a given text, a system that aims to support language learning first needs some objective assessment of the text's complexity or its corresponding operationalization, the text's readability. Text complexity is characterized by the following dimensions:

Lexical The text contains rare or ambiguous words.

Morphological Rare morphological particles (word formation processes) are used. This factor is particularly important for agglutinative languages (e.g. Japanese, Turkish).[2]

Syntactic Complex grammatical structures are used. Advanced syntactic constructions (e.g. embedded sentences) increase the complexity of the text.

Semantic Infrequent senses of words are used or meanings are composed in an unusual way (e.g. idioms).

Discourse The argument structure of the text is not explicitly mentioned.

Conceptual The text requires domain knowledge. Texts about philosophy or math might be stylistically easy, but require extensive conceptual background knowledge.

Pragmatic The interpretation of the text is twisted by the text genre. The content might be understandable, but the author's intention needs advanced interpretation as is often the case in satire.

Readability measures automatically estimate the complexity of a text based on features from several

[2]Agglutinative languages use a high number of affixes to change the meaning of a word.

of the above dimensions. In the following, we group the features according to their dimension and elaborate on their use in readability measures. The dominant language for readability approaches is English. Therefore, we neglect the morphological dimension in the overview.[3] As readability measures are not yet capable of capturing the conceptual and the pragmatic dimension, we also omit them. In general, newer approaches incorporate most of the features from the previous work. Therefore, we only discuss the new features each approach contributes.

2.1 Lexical dimension

Features in the lexical dimension capture the difficulty of the vocabulary of a text. The choice of words has a strong effect on the comprehensibility of a text; this holds especially for language learners.

Surface-based measures The traditional readability measures rely on two main features: word length and sentence length. They are computed by the average number of characters (or syllables) per word and the average number of words per sentence[4], and are combined with manually determined weights resulting in a grade level as output. Most prominent methods of this type are the Flesch–Kincaid Grade Level (Kincaid et al., 1975), the Automatic Readability Index (Smith and Senter, 1967) and the Coleman–Liau Index (Coleman and Liau, 1975). The Fry Formula (Fry, 1977) plots the word length and the sentence length on a graph and defines areas for each grade level. The corresponding grade level for a text and also the distance to neighboring grade levels can then easily be read from the graph. In addition to the word and sentence length, the SMOG grade (McLaughlin, 1969) and the Gunning–Fog Index (Gunning, 1969) also consider the number of complex words defined as words with three or more syllables. Some of these surface-based approaches are employed in standard word processors. However, they have also been subject to criticism as they only capture structural characteristics of the text and can easily be misleading.[5]

[3] In section 3, we summarize readability measures for other languages

[4] As a common pre-condition, the text should usually contain a minimum of 100 words.

[5] See DuBay (2004) for a very detailed overview of the strengths and weaknesses of the surface-based measures.

For English, word length is a very good approximation of difficulty, as frequently used words tend to be rather short compared to more specific terms (Sigurd et al., 2004). However, there exist of course many exceptions to this.[6] Alternatively, the method described by Dale and Chall (1948) proposes the use of word lists that are based on the frequency of words. If many words of a text do not occur in the list, this serves as an indicator for higher text complexity.

Language models Instead of absolute frequencies as in word lists, language model approaches are based on word probabilities. The use of language models is a common technique in speech recognition and machine translation in order to determine the probability of a term in a given context. Collins-Thompson and Callan (2005) have shown that this notion of the probability of a term can easily be transferred to readability, since it is generally assumed that a sentence is more readable if it uses very common terms and term sequences. In combination with smoothing methods and pre-processing (e.g. stemming), language models can also account for novel combinations of words. Higher n-gram models as used by Schwarm and Ostendorf (2005) can even account for collocation frequencies indicating different usages of content words (e.g. *hit the ball* / *hit rock bottom*). Language models can easily be re-trained for new domains and new languages; they are therefore particularly suitable in self-directed learning. They return a probability distribution of terms over all readability levels.

Lexical variation The lexical difficulty of a text is not only determined by the choice of words, but also by the amount of lexical variation. If the same concept is expressed by different words, the reader has to recognize the similarity relation of the words in order to understand the shared reference. Lexical variation is usually measured by the type-token ratio (Graesser and McNamara, 2004), where type is a word and token refers to the different usages of the word in the text. A low ratio indicates that words are frequently repeated in the text. This characteristic might decrease the stylistic elegance of the text, but it facilitates text comprehension.

[6] Compare, for example, *together* (length 8, ANC frequency 4004) and *sag* (length: 3, ANC frequency: 27)

2.2 Syntactic dimension

Syntactic features measure the grammatical difficulty of the text. Especially for language learners, complex syntactic structures are major text comprehension obstacles. The surface-based measures estimate the syntactic difficulty by considering sentence length (see section 2.1). However, although a longer sentence might indicate a more complex structure, it could also simply contain an enumeration of concepts. In recent approaches, the grammatical structure is thus represented by part-of-speech (POS) patterns and parse trees, as described below.

POS tagging In readability measures, POS tagging is mainly used for the distinction of content and function words. Content words carry lexical meaning, while function words like articles or conjunctions indicate syntactic relations. A high number of content words indicates high lexical density (Vajjala and Meurers, 2012). Feng and Huenerfauth (2010) additionally determine the absolute and relative numbers of the different POS tags in the sentence and found that a high number of nouns and prepositions is an indicator for text complexity. Heilman et al. (2007) highlight the occurrence of different verb tenses as indicators for text complexity, especially for second language learners. Grammatical constructions are usually acquired step by step and complex structures such as the use of the passive voice occur in later stages. Infrequent verb tenses might thus strongly inhibit a learner's comprehension of the text.

Parsing In addition to POS information, parsing features are used for predicting readability. Syntactic parsers analyze the grammatical structure of a sentence and return a formal syntax representation. For readability measures, the number and type of noun and verb phrases are determined (Schwarm and Ostendorf, 2005; Heilman et al., 2007). In addition, Schwarm and Ostendorf (2005) include the depth of the parse tree and the number of subordinated sentences in order to model the sentence complexity. Similarly, Vajjala and Meurers (2012) consider the number of clauses per sentence and the number of subordinations and coordinations. Another parsing feature, used by Tonelli et al. (2012), is the syntactic similarity of sentences. A text is easier to read if it exhibits low syntactic variability. This can be computed by detecting the largest common subtree of two sentences. When accessing user profiles for second language learning, it is possible to determine even more concrete syntactic features that decrease the comprehensibility of a text for a specific learner.

2.3 Semantic dimension

The semantic dimension is related to the meaning of words and sentences. Lexical semantics captures the meaning of words, while compositional semantics describes the sentence meaning.

Lexical semantics Polysemous words complicate the interpretation of a sentence because they have to be disambiguated first. Words denoting abstract concepts, on the other hand, are considered difficult because they do not describe a concrete object. In the CohMetrix readability framework (Graesser and McNamara, 2004), polysemy and abstractness are determined on the basis of WordNet relations (Fellbaum, 1998). Polysemy is measured by the number of synsets of a word and abstractness is determined by the number of hypernym relations.

Compositional semantics The semantics of a sentence can be represented by semantic networks consisting of conceptual nodes linked by semantic relations. Vor der Brück et al. (2008) applied the semantic Wocadi-Parser (Hartrumpf, 2003) for their readability measure on German texts. They considered the number of nodes and relations in the semantic representation as indicators of semantic complexity. These features correlate well with human judgments of readability, but the parser often fails to build a representation, limiting the robustness of their approach. The concepts of polysemy and abstractness can be determined more easily.

2.4 Discourse dimension

In the readability literature, all intersentential relations are perceived as discourse related. Discourse features model the structure of the text as indicated through cohesive markers and the coherence of arguments through reference resolution.

Cohesion An important indicator for text cohesion is the use of discourse connectives. Pitler and

Nenkova (2008) build a discourse language model based on the annotations from the Penn Discourse Bank. This model determines how likely it is for each grade level that the text contains implicit or explicit discourse relations. Tonelli et al. (2012) manually create a list of additive, causal, logical, and temporal connectives for Italian. In addition, they capture the "situation model dimensions of the text" by calculating the ratio between causal or intentional particles and causal or intentional verbs. Causal and intentional verbs are identified manually by exploiting category and gloss information from WordNet.

Coherence The coherence of a text can be measured by the pronoun density. If concepts are not named directly, but referenced by a pronoun, the resolution of the meaning is more difficult. Graesser and McNamara (2004) analyze co-references in more detail by determining the relations between two consecutive sentences. Noun overlap and stem overlap in the sentence pair are both indicators for coherence. Alternatively, Pitler and Nenkova (2008) generate entity grids that capture how the center of attention shifts from one entity in the text to another as postulated in the centering theory (Grosz et al., 1995). Feng and Huenerfauth (2010) keep track of the number of entity mentions. Additionally, they assume that a higher number of active entities poses a higher working memory load on the reader. In order to determine the active entities, they identify lexical chains. A lexical chain is formed by entities that are linked through semantic relations such as synonymy or hyponymy. The length and the sentence span of the chain are interpreted as indicators for text complexity.

2.5 Combining features

From a diachronous view, readability measures have continuously taken more and more features into account. Early measures in the 1960s worked only with surface-based features and manually adjusted the parameters. Later approaches successively added features from the lexical, syntactic, semantic, and discourse dimensions as the respective technologies became available. As the number of features was steadily growing, the need for machine learning methods emerged. Supervised learning methods use training data to determine the significant features for each grade level. Using the learned feature weights then enables the prediction of grade levels for unseen texts. A common training corpus contains news articles for educational use from the WeeklyReader[7] that are labeled according to the US grade levels. Several learning algorithms have been applied for readability measures—e.g. Naïve Bayes (Collins-Thompson and Callan, 2005), k-nearest neighbors (Heilman et al., 2007), support vector machines (Schwarm and Ostendorf, 2005) and linear regression (Pitler and Nenkova, 2008). Tanaka-Ishii et al. (2010) used data annotated with only two different reading classes. This enabled the use of a sorting algorithm that sorts texts according to their readability instead of returning an absolute value. Heilman et al. (2008) compare different machine learning approaches that respectively interpret the readability grades as nominal, ordinal and interval scales of measurements. In their setting, interpreting the readability scores as ordinal data performed best. Thus, the scores are considered to have a natural ordering, but they are not evenly spaced.

The use of feature combination for readability measures has become the common approach, but it has not yet been discussed how these need to be adapted to other languages, to language learning, and to self-directed learning.

3 Adaptation to other languages

The applicability of the explored readability features to other languages is poorly studied because most approaches focus on English. Statistical methods such as language models can easily be adapted to other languages, parsers and POS-taggers are not always available in a comparable quality. Several researchers ported the methods that worked successfully for English to other languages. François and Fairon (2012) implement a readability measure for French, and Aluisio et al. (2010) for Portuguese. Tonelli et al. (2012) rely on the CohMetrix framework and implement an Italian version of the features.

However, features established for English are not necessarily significant for languages with different properties. The particular characteristics of a given language should also be considered in the feature se-

[7] http://www.weeklyreader.com/

lection. Collins-Thompson and Callan (2005), for example, come to the conclusion that their language model–based approach heavily benefits from stemming when applied to the more inflected language French. Similarly, Dell'Orletta et al. (2011) introduce morphological features for Italian. Vor der Brück et al. (2008) present a readability measure for German and also rely on extensive morphological analysis. In addition, they add features specific to German such as the distance between a verb and its separable prefix. Larsson (2006) introduce a new feature for Swedish that identifies subordination by the use of the Swedish conjunction *att*. Sato et al. (2008) present a readability measure for Japanese and introduce new features in order to deal with the different character sets. Another problem for Japanese is the detection of word boundaries as they are not indicated by white space. Al-Khalifa and Al-Ajlan (2010) experiment with readability measures for Arabic and address similar issues related to the different character set.

These examples show that readability can be measured by different text characteristics depending on the specific language. More focused research is necessary in order to determine the most predictive features for each language. However, some major features such as lemma frequency are shared across most languages. They can approximate the readability even for under-resourced languages.

4 Adaptation to language learning

The acquisition of the native language (L1) and the process of learning a second language (L2) evolve in different ways. The three main differences are the age of acquisition, the mode of acquisition and the background knowledge (see table 2). Most of the introduced readability measures have been established for native speakers of English, while aspects of foreign language learning have not yet been studied in detail. Vajjala and Meurers (2012) use features that are motivated in the evaluation of language learners' written production. However, it remains unclear how these features differ from those for native speakers.

L2 learner grades The native language is usually acquired in the first years of childhood, while an L2 is generally learned on top of the L1.[8] This means that a certain level of proficiency in the L1 already exists. As learners are older when learning an L2, they also tend to have a more advanced educational background and have already developed higher intellectual abilities (Cook et al., 1979). On the other hand, L2 learning usually progresses significantly slower than the native language acquisition. Due to these differences, school grade levels indicating the readability of L1 texts cannot be directly mapped to foreign language learning, but rather need to be learned individually from L2 data. The readability for L2 texts should thus not be expressed in school grades, but in L2-specific learner levels.

Fine-grained feedback Language acquisition of the native language is a strongly debated topic in psychology and pedagogy. We will not further elaborate on the cognitive aspects of this process. However, one general difference of the learning setting needs to be considered: the basic L1 knowledge is learned from the unstructured input children receive from the environment, while an L2 is usually learned gradually by instruction (Cook et al., 1979). An L2 can also be learned by simple exposure to the language (informal language learning (Bahrani and Sim, 2012)), but it is usually a more conscious process that also requires more structured input (Schmidt, 1995). School children have already acquired the basic structures of their L1, while L2 students need to actively learn new syntactic regularities. This indicates that the output of readability measures has to be more fine-grained than standard school grades.

The evaluation of supervised learning approaches has shown that syntactic features in isolation perform significantly worse than lexical features in predicting the correct school grade for L1 texts (Heilman et al., 2007; Feng and Huenerfauth, 2010). The syntactic features contribute only slightly to the improvement of the overall readability prediction. However, for L2 learners the extensive use of an unknown verb tense can be a stand-alone criterion for the readability of a text. In the feature combination, this individual information might be lost and cannot fully characterize the text complexity. An ap-

[8] Except for bilingual children who acquire two languages simultaneously

propriate readability measure for L2 learning should thus provide more fine-grained information about the readability. As a result, the language learning system receives information about the lexical, syntactic, semantic, and discourse difficulty of the text and can adapt the learning setting accordingly.

Consideration of L1 The L2 learning is influenced by the background knowledge of the learner. As L1 is already present, basic concepts of languages such as the different behavior of word classes or the syntactic coordination of arguments are already known. In addition, the specific properties of the L1 influence the acquisition of the L2. The phenomena of cross-linguistic transfer have been heavily researched (Odlin, 1989; Zobl, 1980). For example, foreign words that have a similar stem as the translation in the mother tongue are acquired more easily. Similarly, syntactic structures that are comparable across the two languages are less error-prone than idiosyncratic aspects of the L2. Thus, readability measures should account for the native language of the learner and should be adapted to groups of users sharing a common mother tongue. The consideration of a learner-specific feature establishes a focus on user profiles which is even more relevant for self-directed learning.

5 Adaptation to self-directed learning

In the setting of self-directed learning, the user profiles can be more heterogeneous than in school classes. The users differ in age, previous knowledge, intellectual ability, and educational and cultural background, and also might have different learning goals. To account for this, a fine-grained learner model is needed, which captures the learner's knowledge and preferences (Al-Hmouz et al., 2010). A model needs to be instantiated based on the learner's knowledge and updated according to the ongoing performance. The previous knowledge can either be estimated by a pre-test or automatically learned from texts that the learner has already mastered. The update function should dynamically assess the performance in exercises and also consider the learner's usage patterns of the system in order to identify preferences for certain exercises. For example, Virvou and Troussas (2011) maintain an error model in order to keep track of the learner's weaknesses.

The learner model needs to be incorporated into the readability measure in order to determine the readability of a text for one specific learner. The readability measure should model the discrepancy between the characteristics of the text (represented by the extracted features) and the learner's knowledge (represented by the learner model). Thus, the measure models not only the general readability of the text, but also its *suitability* for a specific learner.

A personalized language model that represents the learner's lexical knowledge could be directly compared to the lexical features of the text. However, a one-to-one mapping from knowledge representation to features is not always possible. For example, if the learner has a recorded preference for sports texts, this translates into features from several dimensions (i.e. advanced sports vocabulary, preference for factual style, acquaintance with sports entities, and domain knowledge). As an approximation, the readability measure could assign a degree of difficulty to each dimension. Each dimension can then be looked up in the learner model to verify the competence level of the learner. The suitability of a text for a specific learner could then be expressed by the discrepancy between the learner competence and the text characteristics for each dimension. This allows more fine-grained support for the text elements that cause difficulties for the learner.

6 Application

In an adaptive language learning system, automatic exercise generation plays an important role in accounting for the variability of learners. A precondition for useful automatic exercise generation is a readability measure that gives fine-grained information about the suitability of a text for a certain learner.

Generating suitable exercises for language learning can be approached from two perspectives: it can either be input-driven or determined by a curriculum. The input-driven method utilizes the learner's interests and is embedded into her routines. The learner can select a text in the foreign language that appears particularly interesting or that needs to be read anyway. The system then generates questions on the basis of the text (bottom-up) in order to fa-

cilitate comprehension and to assist with unknown words or constructions.

In the curriculum method, the learning goal is predefined by a learning framework (i.e. realizations of the learner levels as defined by the *Common European Framework of Reference for Languages*[9]). The learner is supposed to learn a new concept (e.g. a grammatical phenomenon, a group of related words) and the exercises are generated in order to reach this goal (top-down). In addition to the learning goal, the exercises should also consider the previous knowledge of the student. A text that meets the learner's interests and knowledge level better stimulates the intrinsic motivation to learn.

For input-driven scenarios, the readability measure can help to extract the dimension of the text that causes comprehension difficulties and trigger exercises to resolve them. The exercise type and the exercise difficulty will thus be determined by the readability outcome—e.g. low readability in the lexical dimension triggers vocabulary exercises. In the case of a given learning goal, the measure helps to acquire the most suitable reading material that best matches the user's profile and fulfills the requirements of the learning goal.

7 Conclusions

In this paper, we gave an overview of readability measures from the perspective of self-directed language learning. We discussed how readability measures need to be adapted in order to consider the requirements of other languages, the different progress levels in L2 acquisition, and the characteristics of user profiles. We suggest the introduction of L2 learner grades and a more fine-grained level of readability feedback. In addition, we propose to assess the suitability of a text with respect to a user model. In the future, we will further develop and implement the proposed measures, and apply them for automatic exercise generation.

Acknowledgments

This work has been supported by the Volkswagen Foundation as part of the Lichtenberg-Professorship Program under grant No. I/82806, and by the Klaus Tschira Foundation under project No. 00.133.2008.

[9] http://www.coe.int/t/dg4/linguistic/Cadre1_en.asp

References

Ahmed Al-Hmouz, Jun Shen, Jun Yan, and Rami Al-Hmouz. 2010. Enhanced learner model for adaptive mobile learning. In *12th International Conference on Information Integration and Web-based Applications & Services - iiWAS '10*, pages 783–786, New York, USA. ACM Press.

Hend S. Al-Khalifa and Amany Al-Ajlan. 2010. Automatic Readability Measurements of the Arabic Text: An Exploratory Study. *The Arabian Journal for Science and Engineering*, 35(2C).

Sandra Aluisio, Lucia Specia, Caroline Gasperin, and Carolina Scarton. 2010. Readability assessment for text simplification. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–9. Association for Computational Linguistics, June.

Taher Bahrani and Tam Shu Sim. 2012. Informal language learning setting: technology or social interaction? *The Turkish Online Journal of Educational Technology*, 11(2):142–149.

Rebekah George Benjamin. 2011. Reconstructing Readability: Recent Developments and Recommendations in the Analysis of Text Difficulty. *Educational Psychology Review*, 24(1):63–88, October.

Mery Coleman and TL Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283–284.

Kevyn Collins-Thompson and Jamie Callan. 2005. Predicting Reading Difficulty with Statistical Language Models. *Journal of the American Society for Information Science and Technology*, 56(13):1448–1462.

Vivian J. Cook, John Long, and Steve McDonough. 1979. First and second language learning. *The Mother Tongue and Other Languages in Education*, pages 7–22.

Edgar Dale and Jeanne S Chall. 1948. A formula for predicting readability: Instructions. *Educational Research Bulletin*.

Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2011. READ-IT: assessing readability of Italian texts with a view to text simplification. In *Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies*, pages 73–83, July.

William H. DuBay. 2004. The Principles of Readability. *Impact Information*, pages 1–76.

Christiane Fellbaum. 1998. *WordNet: An electronic database*. MIT Press, Cambridge, MA.

Lijun Feng and Matt Huenerfauth. 2010. A Comparison of Features for Automatic Readability Assessment. In *Proceedings of COLING 2010*, pages 276–284, August.

Thomas François and Cédrick Fairon. 2012. An AI readability formula for French as a foreign language. In *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 466–477, July.

Edgar Fry. 1977. Fry's readability graph: Clarifications, validity, and extension to level 17. *Journal of Reading*.

Arthur C. Graesser and Danielle McNamara. 2004. CohMetrix: Analysis of text on cohesion and language. *Behavior Research Methods*, 36(2).

Barbara J. Grosz, Scott Weinstein, and Aravind K. Joshi. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational linguistics*, 21(2):203–225.

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

Sven Hartrumpf. 2003. *Hybrid Disambiguation in Natural Language Analysis*. Der Andere Verlag, Osnabrück, Germany.

Michael J 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 *Proceedings of NAACL-HLT*, pages 460–467.

Michael Heilman, Kevyn Collins-Thompson, and Maxine Eskenazi. 2008. An Analysis of Statistical Models and Features for Reading Difficulty Prediction. In *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications - EANL '08*, pages 71–79, Morristown, NJ, USA, June.

Filip Karel and Jií Klema. 2006. Adaptivity in e-learning. *Current Developments in Technology-Assisted Education*, 1:260–264.

John P. Kincaid, Robert P. Fishburne Jr, Richard L. Rogers, and Brad S. Chissom. 1975. Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel. Technical report, DTIC Document.

Patrik Larsson. 2006. *Classification into Readability Levels Implementation and Evaluation*. Master's thesis, Uppsala University, Sweden.

G. Harry McLaughlin. 1969. SMOG grading: A new readability formula. *Journal of reading*, 12(8):639–646.

Terence Odlin. 1989. *Language Transfer: Cross-Linguistic Influence in Language Learning*. Cambridge University Press.

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*, pages 186–195. Association for Computational Linguistics.

Satoshi Sato, Suguru Matsuyoshi, and Yohsuke Kondoh. 2008. Automatic Assessment of Japanese Text Readability Based on a Textbook Corpus. In *6th International Conference on Language Resources and Evaluation (LREC)*, Marrakech, Morocco.

Richard Schmidt. 1995. Consciousness and foreign language learning: A tutorial on the role of attention and awareness in learning. *Attention and Awareness in Foreign Language Learning*, pages 1–63.

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 of the ACL*, pages 523–530, June.

Bengt Sigurd, Mats Eeg-Olofsson, and Joost van Weijer. 2004. Word length, sentence length and frequency - Zipf revisited. *Studia Linguistica*, 58(1):37–52, April.

E. A. Smith and R.J. Senter. 1967. *Automated readability index.* Cincinnati University Ohio.

Kumiko Tanaka-Ishii, Satoshi Tezuka, and Hiroshi Terada. 2010. Sorting Texts by Readability. *Computational Linguistics*, 36(2):203–227, June.

Sara Tonelli, Ke Tran Manh, and Emanuele Pianta. 2012. Making Readability Indices Readable. In *Proceedings of NAACL-HLT: Workshop on Predicting and Improving Text Readability for target reader populations*, pages 40–48.

Sowmya Vajjala and Detmar Meurers. 2012. On Improving the Accuracy of Readability Classification using Insights from Second Language Acquisition. In *Proceedings of the 7th Workshop on Innovative Use of NLP for Building Educational Applications (BEA7)*, pages 163–173.

Maria Virvou and Christos Troussas. 2011. Web-based student modeling for learning multiple languages. In *International Conference on Information Society (i-Society)*, pages 423–428.

Tim Vor der Brück, Sven Hartrumpf, and Hermann Helbig. 2008. A Readability Checker with Supervised Learning using Deep Syntactic and Semantic Indicators. In *11th International Multiconference: Information Society-IS*, pages 92–97.

Lev Vygotsky. 1978. Mind in society: The development of higher psychological processes. Harvard University Press.

Helmut Zobl. 1980. Developmental and transfer errors: their common bases and (possibly) differential effects on subsequent learning. *TESOL Quarterly*.