Comparing Machine Learning Classification Approaches for Predicting Expository Text Difficulty

Renu Balyan, Kathryn S. McCarthy, Danielle S. McNamara

Published May 2018

The Thirty-First International Florida
Artificial Intelligence Research Society Conference (FLAIRS-31)

Comparing Machine Learning Classification Approaches for Predicting Expository Text Difficulty

Renu Balyan, Kathryn S. McCarthy, Danielle S. McNamara

Arizona State University, Tempe, AZ, USA
{renu.balyan; ksmccar1; dsmcnamara}@asu.edu

Abstract

While hierarchical machine learning approaches have been used to classify texts into different content areas, this approach has, to our knowledge, not been used in the automated assessment of text difficulty. This study compared the accuracy of four classification machine learning approaches (flat, one-vs-one, one-vs-all, and hierarchical) using natural language processing features in predicting human ratings of text difficulty for two sets of texts. The hierarchical classification was the most accurate for the two text sets considered individually (Set A, 77.78%; Set B, 82.05%), while the non-hierarchical approaches, one-vs-one and one-vs-all, performed similar to the hierarchical classification for the combined set (71.43%). These findings suggest both promise and limitations for applying hierarchical approaches to text difficulty classification. It may be beneficial to apply a recursive top-down approach to discriminate the subsets of classes that are at the top of the hierarchy and less related, and then further separate the classes into subsets that may be more similar to one other. These results also suggest that a single approach may not always work for all types of datasets and that it is important to evaluate which machine learning approach and algorithm works best for particular datasets. The authors encourage more work in this area to help suggest which types of algorithms work best as a function of the type of dataset.

Introduction

Gaining new knowledge, in both formal and informal environments, relies heavily on learning from text. An important component of the comprehension process is the difficulty of the text being read. Texts that are too difficult can impede comprehension. Educators find texts that are grade appropriate and may also need to select texts that meet the need of individual student. Given the abundance of text materials available, educators simply do not have the time to find and thoroughly evaluate texts for this purpose. As such, educators depend on text difficulty formulas to quickly identify appropriately challenging texts. Common readability formulas such as Flesch-Kincaid Reading Ease (Flesch, 1948) that assess text difficulty are usually based on number of syllables per word, number of words per sentence, and the number of sentences (Klare, 1974). Though easy to use, these formulas are centered on relatively shallow lexical and sentential indices. However, theories of reading comprehension suggest that deep features related to syntax and semantics drive text difficulty (Dufty et al., 2006; Duran et al., 2007; McNamara, Graesser, and Louwerse, 2012). To address this issue, researchers have developed natural language processing (NLP) tools that extract richer information about the linguistic features of a text that reflect complex dimensions such as narrativity, syntactic complexity, and cohesion (e.g. Crossley et al., 2016; McNamara, et al., 2014).

Researchers have also begun to employ machine-learning approaches for measuring text readability (Collins-Thompson, 2014; Kate et al., 2010; Kotani, Yoshimi, and Isahara, 2011; Pilán, Volodina, and Johansson, 2014). These approaches have shown promise in more accurately assessing text difficulty as compared to "classic" readability approaches (François and Miltsakaki, 2012).

Though promising, most of this work has focused on either determining the best set of linguistic features or comparing regression and classification approaches (e.g., François and Miltsakaki, 2012; Heilman et al., 2008). To our knowledge, there is little work investigating the potential for hierarchical approaches in the classification of text difficulty. Hierarchical classification has been used in a number of areas such as protein classification (Zimek et al., 2008), essays scoring (McNamara et al., 2015), and automatic target recognition (Casasent and Wang, 2005). This study addresses this gap in the literature by combining NLP and machine learning to compare multiple types of classification in their accuracy of classifying text difficulty.

We first provide brief description of the relevant NLP tools and machine learning techniques and then present results of the experiments.

Natural Language Processing

LP intersects computational linguistics, computer science, and artificial intelligence to understand, assess, and respond to naturally occurring human language. NLP has been used in education to support student learning, for intelligent and automatic assessments, to improve learning and teaching in massive open online courses (MOOCs), and to develop learning systems. In this study, we employed the NLP tool, Coh-Metrix (McNamara et al., 2014), which integrates a number of sophisticated tools such as advanced syntactic parsers, part-of-speech taggers, distributional models, and psycholinguistic databases (Coltheart, 1981) to generate over 400 indices of language, text, and readability.

Machine Learning

Machine learning algorithms are categorized as unsupervised and supervised. Unsupervised learning uses data that is not labeled, whereas in supervised machine learning, the algorithms are trained on labeled data. For supervised algorithms, *regression* is used to predict quantitative variables, whereas *classification* is used to predict qualitative variables (Hastie, Tibshirani, and Friedman, 2009; James et al., 2013). As our data involves human ratings of categories (i.e. labeled categorical data), we adopted a supervised learning classification approach.

Commonly used classification algorithms include Decision Trees, Naïve Bayes, Linear Discriminant Analysis (LDA), Support Vector Machines (SVM), Logistic Regression, Random Forests, Neural Networks, and Boosting (for further description of these algorithms, see Balyan, McCarthy, and McNamara, 2017; Hastie et al., 2009). Preliminary experiments with a number of these algorithms indicated that SVM and LDA were the most accurate for the current data.

Non-Hierarchical Classification

Three types of non-hierarchical approaches were used in this study. *Flat classification* is the simplest and the most direct approach to category classification. It uses single classifier and all class variable instances in the training dataset. The other two non-hierarchical approaches: *one-vs-one* and *one-vs-all* use SVM, which relies on the construction of multiple hyperplanes. SVMs are typically designed for binary classification (Duda, Hart, and Stork, 2000). However, *one-vs-one* and *one-vs-all* are the two most popular approaches for extending SVMs to classify K-classes (K>2). The one-vs-one classification approach forms a binary classifier for each class-pair and hence k*(k-1)/2 classifiers are required. In contrast, one-vs-all classification compares each class to all other classes (Duda et al., 2000; James et al., 2013).

Hierarchical Classification

The one-vs-all approach does not consider similarity between the classes (Casasent and Wang, 2005). Additionally, one-vs-one approach is not attractive when K is large (Kumar et al., 2002). A more preferred approach is a *binary hierarchical classification* (Wang and Casasent, 2009). This approach is based on the divide-and-conquer strategy, and learns concepts more effectively and efficiently (Kumar and Ghosh, 1999). In some cases, the hierarchical classifier has performed better than some single complex classifiers, such as neural networks (NNs) and kNN classifiers (Kumar, Gosh, and Crawford, 2002; Schwenker, 2000). In binary hierarchical classification, the classes are divided into two smaller macro-classes at each node. Only $\log_2 K$ classifiers need to be traversed in order to move from the top to a bottom decision node.

Current Study

This study leverages NLP to compare machine learning classification approaches in their accuracy of classifying human ratings of text difficulty. We conducted these experiments in context of text development for our reading comprehension intelligent tutor, Interactive Strategy Training for Active Reading and Thinking (iSTART; McNamara, Levinstein, and Boothum, 2004). This study is not only theoretically interesting, but also has important applications. By using accurate NLP tools and machine learning algorithms, we can automate the process of classifying new texts within the iSTART text library. If successful, the current approach will permit researchers and teachers to add their own texts into the text library ensuring that the difficulty levels assigned in the system remain consistent over time.

We used two text sets, both individually and combined, to compare not only different classifiers, but also four different approaches. The first set of experiments identified the most accurate classification algorithms (SVM and LDA) and the second set of experiments established which approach (flat, one-vs-one, one-vs-all, and hierarchical) was most effective for classifying human ratings of text difficulty.

Method

Corpus

The text corpus was comprised of two text sets developed for iSTART, an intelligent tutoring system (ITS) that supports successful reading comprehension of complex informational texts through self-explanation training (McNamara, Levinstein, and Boonthum, 2004; Snow et al., 2016).

Set A was comprised of texts from the iSTART StairStepper module (n = 162), including expository texts, ranging in topics from science, history, pop culture, and

sports. These texts were collected from reading comprehension tests culled from publicly available websites (Perret, et al., 2017). Set B was comprised of texts from iSTART's main text library (n = 100), which are used for self-explanation practice and within various games. These texts are complex, informational texts about scientific phenomena compiled when developing the practice modules within iSTART (Jackson and McNamara, 2013). The texts were culled from various sources, primarily science textbooks. Whereas Set A included texts that varied widely in genre and difficulty (i.e., grades 1-12), Set B comprised texts used to provide information typical in high school and college science courses. Hence, the two sets of texts were quite different in nature.

The difficulty of the texts in the two sets were rated separately, but followed the same procedure (see Johnson, et al., 2017). Set A was sorted into 12 levels of difficulty and Set B was sorted into 9 levels. We compared the levels across the two sets and determined that the easiest texts in Set B were of equal difficulty to Set A texts rated as difficulty level 6. Thus, combining Set A (1-12) and Set B (6-14) texts resulted in a corpus that included 262 texts categorized into 14 difficulty levels.

Initial machine learning experiments that considered all text difficulty levels (1-12 for Set A and 6-14 for Set B) resulted in low accuracy: 25.97% to 33.95% for Set A and 29.17% to 35.42% for Set B. The accuracy decreased further when the two sets were combined: ranging from 19.44% to 26.39%. Consequently, we clustered the 14 levels into more coarse-grained levels. The researchers re-read the texts in each difficulty level to identify intuitive breaks in the text set. This resulted in four difficulty levels: low (1-4), middle (5-8), high (9-12), and very high (13 and 14). Set A included low, middle, and high difficulty texts, while Set B included middle, high, and very high difficulty texts. These four levels are roughly aligned with levels of schooling in the United States: elementary (1-4), middle school (5-8), high school (9-12), and college-appropriate (13 and 14).

Selection of Linguistic Indices

We selected 11 linguist indices related to lexical sophistication, readability, lexical diversity, and syntactic complexity that have been shown to correlate with text difficulty (e.g., Crossley, Allen, and McNamara, 2012; Salsbury, Crossley, and McNamara, 2011). Removing highly correlated indices (Pearson's r > .85) reduced this to eight indices. The following sections provide brief descriptions of each of these indices used for the experiments.

Flesch-Kincaid Grade Level

Flesch-Kincaid Grade Level (FKGL; Kincaid et al., 1975) is a simple measure of readability computed using average syllables per word (ASW) and average sentence length (ASL). Lower grade levels correspond to easier texts.

L2 Readability

L2 readability score predicts the readability of texts for second-language learners (Crossley, Allen, and McNamara, 2012). L2 readability score considers content word overlap, sentence syntactic similarity, and word frequency. In contrast to FKGL, higher scores indicate easier texts.

Syntactic Complexity

Syntactic complexity of a sentence is determined by considering mean number of words before the main verb and higher number of higher-level constituents per word in the sentence. Sentences with less syntactic complexity are easier to process and comprehend (Crossley, Allen, and McNamara, 2012; Perfetti, Landi, and Oakhill, 2005).

Uncommon or Rare Words

Uncommon or rare word indices are indicative of the frequency that a word occurs in the English language. More uncommon or rare words in a text render the text more difficult. The text difficulty is expected to increase if there are words that readers have never or rarely encountered. This index is computed from CELEX (Baayen, Piepenbrock, and Gulikers, 1995), a 17.9 million words corpus.

Lexical Diversity

Lexical diversity, or the variety of words used in a text, is often measured using type–token ratios (TTR). However, this measure is highly correlated to text length. In order to assess lexical diversity regardless of text length, we employed MTLD (measure of textual, lexical diversity; McCarthy, 2005) and D values (Malvern et al., 2004; McNamara, Crossley, and Roscoe, 2013).

Word Familiarity

Sentences that contain words that are more familiar are processed more quickly (McNamara et al., 2013). Word familiarity is a human rating of how easily adults recognize a word. For example, the word 'dog' has a higher average familiarity than 'cortex'. Word familiarity ratings are computed using the MRC Psycholinguistic Database, which provides ratings for several thousand words along several psychological dimensions.

Word Imagability

Word imagability refers to the ease with which one can construct a mental image of a word in one's mind. For example, 'airplane' and 'hammer', are more imagable than 'dogma' and 'quantum'.

Table 1: Means and ANOVA Results for the selected linguistic feature

	Set A				Set B				A + B				
Feature	**Low**	**Middle**	**High**	**F (2,159)**	**Middle**	**High**	**Very High**	**F (2,97)**	**Low**	**Middle**	**High**	**Very High**	**F (3,258)**
FKGL	5.87	8.83	10.83	113.40	9.17	9.25	11.56	17.03	5.87	8.93	10.08	11.56	76.78
L2 Reada-bility	20.64	14.41	12.46	46.17	18.12	17.25	14.34	4.35	20.64	15.47	14.72	14.34	16.80
Syntactic Complexity	0.73	0.68	0.69	13.18	0.72	0.71	0.68	15.71	0.73	0.70	0.70	0.68	11.53
Uncommon /rare Words	49.52	94.87	107.41	21.37	57.92	62.42	81.91	8.39	49.52	84.25	86.15	81.91	8.64
Lexical Diversity (MTLD)	66.41	78.46	81.06	9.69	56.41	57.32	57.15	0.03	66.41	72.12	69.84	57.15	3.92
Age of Acquisition (AoA)	5.13	5.53	5.82	128.40	5.69	5.97	6.39	25.07	5.13	5.57	5.89	6.39	130.00
Imagability	355.11	347.74	336.18	22.46	333.54	329.06	317.99	5.78	355.11	343.66	332.81	317.99	36.58
Familiarity	588.88	587.25	585.71	4.53	588.29	588.51	588.21	0.04	588.88	587.55	587.03	588.21	1.45

Age of Acquisition

Age of Acquisition is an MRC Psycholinguistic Database index that refers to the age at which a word first appears in a child's vocabulary.

A series of analysis of variance (ANOVA) were conducted to determine if each linguistic feature differed as a function of the human ratings of text difficulty for the three different datasets (Set A, Set B, and the combined set referred to A+B). Means and ANOVA results are shown in Table 1.

In the experiments, the omnibus ANOVAs indicated a significant effect of text difficulty level for all indices except *lexical diversity* for Set B and *familiarity* for Sets B and A+B. Post hoc tests revealed that these effects were driven by differences between the low level compared to the high and very high difficulty levels. Few of the indices showed differences between the middle and high levels. These tests as a whole, however, confirmed that the linguistic features vary significantly across text difficulty levels. Consequently, we used these linguistic features in the subsequent machine learning experiments.

Results

Non-Hierarchical Classification

We used ten-fold cross validation to assess the accuracy of the models. The accuracy of the non-hierarchical classification approaches are shown in Table 2. The highest accuracy for the flat classification approach was for Set A (76.19%) and the lowest accuracy was for the A+B dataset (59.47%). LDA classifier achieved the highest accuracy for all the datasets for this approach.

Because Set A and Set B each contained three text difficulty levels (classes), we constructed three classifiers for the one-vs-one and one-vs-all classification tasks. The A+B dataset contained four levels, so we constructed six classifiers for the one-vs-one approach and four classifiers for the one-vs-all approach. SVM was used to train the models. The bold entries in Table 2 indicate highest accuracy for each dataset.

Table 2: Accuracy for non-hierarchical classification

Approach	Classifier	Data Source		
		Set A	Set B	A+B
Flat	LDA	**76.19***	71.79	59.47
one-vs-one	SVM	**76.19***	71.79	**71.43***
one-vs-all	SVM	74.60	**74.36***	**71.43***
* Highest Accuracy				

For Set A, the one-vs-one approach and flat were the most accurate. In contrast, for the Set B the one-vs-all was the most accurate. The one-vs-one and one-vs-all performed similarly for A+B dataset.

Hierarchical Classification

For hierarchical classification, we conducted three experiments with multiple runs. For example, in the first run of Set A, we first classified texts into two classes as 'low' and 'other'. At the second level, the 'other' class was further classified into 'middle' and 'high'. For the second run, the texts were first classified as 'middle' and 'other' and then the 'other' texts were further classified as 'low' and 'high'. Finally, for the third run, the texts were first classified as

'high' and 'other' and then the 'other' texts were reclassified as 'low' and 'middle'. A summary of these experiments for different data combinations is provided in Table 3.

Table 3: Hierarchical Classification Experiments Summary

Experiment	Set A	Set B	A+B
Run 1	L + (M/H)	M + (H/VH)	(L/M) + (H/VH)
Run 2	M + (L/H)	H + (M/VH)	(L/H) + (M/VH)
Run 3	H + (L/M)	VH + (M/H)	(L/VH) + (M/H)
L: Low, M: Middle, H: High, VH: Very High			

The classification accuracy of the final model for all the three experiments is summarized in Table 4. We observed that hierarchical classification (Run 1) improved the accuracy of the model for Set B and the A+B dataset significantly compared with the previous approach (flat classification). In contrast, there was only slight improvement for Set A (Run 3) over the previous approaches.

Table 4: Accuracy for Hierarchical Classification

Experiment	Classifier	Data Source		
		Set A	Set B	A+B
Run 1	LDA and SVM	74.60	**82.05***	**71.43***
Run 2	LDA and SVM	76.19	58.97	66.23
Run 3	LDA and SVM	**77.78***	57.89	70.13
* Highest accuracy				

In sum, we found that hierarchical classification achieved the highest accuracy for Set A (77.78%) and Set B (82.05%). For A+B dataset, we achieved the highest accuracy (71.43%) using both hierarchical as well as the two non-hierarchical approaches, one-vs-one and one-vs-all.

Discussion

This study compared different supervised machine learning approaches in classifying human ratings of text difficulty Our approach was novel in that we submitted the texts to multiple types of machine learning approaches: flat, one-vs-one, one-vs-all, and hierarchical. These experiments demonstrated the potential of using hierarchical approaches in text difficulty classification, but also indicated that no one single approach was most accurate.

When classifying Set A and Set B text sets independently, the most accurate approach was hierarchical classification. However, when the two text sets were combined, one-vs-one, one-vs-all, and hierarchical approaches performed similarly. The differences in the accuracy of these approaches suggest that there are potential differences in the nature of the texts in each text set. As seen in Table 1, the differences between the middle and high text sets were lessened when the sets were combined. Set B texts are scientific texts appropriate for high school and college students, whereas Set A texts were designed to include a broader range of reading skills and topics. Given that the two sets were developed for different purposes and rated independently of one another, it was expected that they would not be perfectly comparable.

At an applied level, we plan to implement separate algorithms to classify text difficulty depending on whether the text will be included in Set A or in Set B within iSTART. This automated classification allows us to continue to permit teachers to add their own texts to the system, while providing an adaptive environment in which students are presented with skill-level appropriate readings.

Acknowledgments

This research was supported in part by the Institute of Education Sciences (IES R305A130124) and the Office of Naval Research (ONR 00014-17-1-2300; ONR N00014-14-1-0343). Opinions or conclusions are those of the authors and do not represent the views of the IES or ONR.

References

Baayen, R. H., Piepenbrock, R., and Gulikers, L. 1995. CELEX. Philadelphia, PA: Linguistic Data Consortium.

Balyan, R., McCarthy, K. S., and McNamara, D. S. 2017. Combining Machine Learning and Natural Language Processing to Assess Literary Text Comprehension. In A. Hershkovitz & L. Paquette (Eds.). In *Proceedings of the 10th International Conference on Educational Data Mining* (EDM), Wuhan, China: International Educational Data Mining Society.

Casasent, D., and Wang, Y-C. F. 2005. A hierarchical classifier using new support vector machine for automatic target recognition. *Neural Networks* 18, 5-6, 541-548.

Collins-Thompson, K. 2014. Computational Assessment of Text Readability: A Survey of Current and Future Research. *ITL - International Journal of Applied Linguistics*, 165(2):97–135.

Coltheart, M. 1981. The MRC Psycholinguistic Database. *Quarterly Journal of Experimental Psychology,* A 33, 4, 497–505.

Crossley, S. A., Allen, D., and McNamara, D. S. 2012. Text simplification and comprehensible input: A case for an intuitive approach. *Language Teaching Research*, 16, 89-108.

Crossley, S. A., Allen, L. K., Snow, E. L., and McNamara, D. S. 2016. Incorporating learning characteristics into automatic essay scoring models: What individual differences and linguistic features tell us about writing quality. *Journal of Educational Data Mining*, 8, 2, 1-19.

Duda, R. O., Hart, P. E., and Stork, D. G. 2000. *Pattern Classification*. Second Edition. Wiley-interscience, New York.

Dufty, D. F., Graesser, A. C., Louwerse, M., and McNamara, D. S. 2006. Assigning Grade Levels to Textbooks: Is it just Readability? In *Proceedings of the 28th Annual Conference of the Cog-*

nitive Science Society, Austin, TX: Cognitive Science Society. In R. Sun and N. Miyake, Eds. 1251-1256.

Duran, N., Bellissens, C., Taylor, R., and McNamara, D. S. 2007. Quantifying Text Difficulty with Automated Indices of Cohesion and Semantics. In *Proceedings of the 29th Annual Meeting of the Cognitive Science Society*. D.S. McNamara and G. Trafton, Eds., Cognitive Science Society, Austin, TX, 233-238.

Flesch, R. F. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32, 221-223.

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

Hastie, T., Tibshirani, R., and Friedman, J. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York.

Heilman, M., Collins-Thompson, K. and Eskenazi, M. 2008. An analysis of statistical models and features for reading difficulty prediction. In *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics, Columbus, OH, USA, 71-79.

Jackson, G. T., and McNamara, D. S. 2013. Motivation and Performance in a Game-Based Intelligent Tutoring System. *Journal of Educational Psychology*, 105, 1036-1049.

James, G., Witten, D., Hastie, T., and Tibshirani, R. 2013. *An Introduction to Statistical Learning with Applications in R*. Springer, New York.

Johnson, A., McCarthy, K., Kopp, K., Perret, C. A., and McNamara, D. S. 2017. Adaptive Reading and Writing Instruction in iSTART and W-Pal. In *Proceedings of the 30th Florida Artificial Intelligence Research Society International Conference* (FLAIRS). AAAI Press.

Kate, R. J., Luo, X., Patwardhan, S., Franz, M., Florian, R., Mooney, R. J., Roukos, S., and Welty, C. 2010. Learning to Predict Readability using Diverse Linguistic Features. In *Proceedings of the 23rd International Conference on Computational Linguistics* (COLING '10). Association for Computational Linguistics, Stroudsburg, PA, USA, 546-554.

Kincaid, J. P., Fishburne, R. P., Rogers, R. L., and 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, Millington, TN: Naval Technical Training, U. S. Naval Air Station, Memphis, TN.

Klare, G. R. 1974. Assessing Readability. *Reading Research Quarterly*, 10 (1), 62-102.

Kotani, K., Yoshimi, T., and Isahara, H. 2011. A Machine Learning Approach to Measurement of Text Readability for EFL Learners Using Various Linguistic Features. US-China Education Review B 6, 767-777.

Kumar, S., Ghosh, J., and Crawford, M. M. 2002. Hierarchical Fusion of Multiple Classifiers for Hyperspectral Data Analysis. Pattern Analysis and Applications, *Spl. Issue on Fusion of Multiple Classifiers*, 5, 2, 210-220.

Kumar, S., and Ghosh, J. 1999. GAMLS: A generalized framework for associative modular learning systems. In *Proceedings of SPIE conference on applications and science of computational intelligence II*, SPIE Proceedings, Orlando, FL, 3722, 24–35.

Malvern, D. D., Richards, B. J., Chipere, N., and Durán, P. 2004. *Lexical Diversity and Language Development: Quantification and Assessment*. Houndmills: Palgrave Macmillan.

McCarthy, P.M. 2005. An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD). Dissertation Abstracts International, 66, UMI No. 3199485.

McNamara, D. S., Crossley, S. A., Roscoe, R. D., Allen, L. K., and Dai, J. 2015. A hierarchical classification approach to automated essay scoring. *Assessing Writing*, 23, 35-59.

McNamara, D. S., Graesser, A. C., McCarthy, P., and Cai, Z. 2014. *Automated evaluation of text and discourse with CohMetrix*. Cambridge University Press, Cambridge.

McNamara, D. S., Crossley, S. A., and Roscoe, R. D. 2013. Natural language processing in an intelligent writing strategy tutoring system. *Behavior Research Methods*, 45, 499-515.

McNamara, D. S., Graesser, A. C., and Louwerse, M. M. 2012. Sources of Text Difficulty: Across Genres and Grades. In *Measuring up: Advances in how we assess reading ability*, J.P. Sabatini, E. Albro, and T. O'Reilly, Eds., Lanham, MD: RandL Education, 89-116.

McNamara, D. S., Levinstein, I. B., and Boonthum, C. 2004. iSTART: Interactive strategy trainer for active reading and thinking. *Behavioral Research Methods, Instruments, & Computers*, 36, 222-233.

Perfetti, C.A., Landi, N., and Oakhill, J. 2005. The Acquisition of Reading Comprehension Skill. In: *The Science of Reading: A Handbook*. M.J. Snowling and C. Hulme Eds. Blackwell, Oxford, 227–247.

Perret, C. A., Johnson, A. M., McCarthy, K. S., Guerrero, T. A., and McNamara, D. S. 2017. StairStepper: An Adaptive Remedial iSTART Module. In B. Boulay, R. Baker & E. Andre (Eds.), *Proceedings of the 18th International Conference on Artificial Intelligence in Education* (AIED), Wuhan, China: Springer. 557-560.

Pilán, I., Volodina, E., and Johansson, R. 2014. Rule-based and machine learning approaches for second language sentence-level readability. In *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*, Baltimore, Maryland USA, 174–184.

Salsbury, T., Crossley, S. A., and McNamara, D. S. 2011. Psycholinguistic word information in second language oral discourse. *Second Language Research*, 27, 343–60.

Schwenker, F. 2000. Hierarchical Support Vector Machines for Multi-Class Pattern Recognition. In *Proceedings of fourth International Conference on knowledge-Based Intelligent Engineering Systems and Allied Technologies*, Brighton, UK, 2, 561–565.

Snow, E. L., Jacovina, M. E., Jackson, G. T., and McNamara, D. S. 2016. iSTART-2: A reading comprehension and strategy instruction tutor. In *Adaptive educational technologies for literacy instruction*, D.S. McNamara and S. A. Crossley, Eds., Taylor and Francis, Routledge: NY, 104-121.

Wang, Y.-C.F., and Casasent, D. 2009. A support vector hierarchical method for multi-class classification and rejection. In *Proceedings of International Joint Conference on Neural Networks*, Atlanta, Georgia, USA, June 14-19, 3281-3288.

Zimek, A., Buchwald, F., Frank, E., and Kramer, S. 2008. A Study of Hierarchical and Flat Classification of Proteins. *IEEE Transactions on Computational Biology and Bioinformatics*, 7 (3). 563-571.