ReaderBench Learns Dutch: Building a Comprehensive Automated Essay Scoring System for Dutch Language

Mihai Dascalu[1,2(✉)], Wim Westera[3], Stefan Ruseti[1], Stefan Trausan-Matu[1,2], and Hub Kurvers[3]

[1] Faculty of Automatic Control and Computers, University "Politehnica" of Bucharest, 313 Splaiul Independenţei, 60042 Bucharest, Romania
{mihai.dascalu, stefan.ruseti, stefan.trausan}@cs.pub.ro
[2] Academy of Romanian Scientists, Splaiul Independenţei 54, 050094 Bucharest, Romania
[3] Open University of the Netherlands, Heerlen, The Netherlands
{wim.westera, hub.kurvers}@ou.nl

Abstract. Automated Essay Scoring has gained a wider applicability and usage with the integration of advanced Natural Language Processing techniques which enabled in-depth analyses of discourse in order capture the specificities of written texts. In this paper, we introduce a novel Automatic Essay Scoring method for Dutch language, built within the *Readerbench* framework, which encompasses a wide range of textual complexity indices, as well as an automated segmentation approach. Our method was evaluated on a corpus of 173 technical reports automatically split into sections and subsections, thus forming a hierarchical structure on which textual complexity indices were subsequently applied. The stepwise regression model explained 30.5% of the variance in students' scores, while a Discriminant Function Analysis predicted with substantial accuracy (75.1%) whether they are high or low performance students.

Keywords: Automated Essay Scoring · Textual complexity assessment · Academic performance · *ReaderBench* framework · Dutch semantic models

1 Introduction

Automated Essay Scoring (AES) is one of the important benefits of Natural Language Processing (NLP) in assisting teachers. AES may analyze the degree to which a student covers in the written text the concepts acquired within the learning process. In addition, it should analyze also the quality of the text, that means its coherence and complexity. Latent Semantic Analysis (LSA) [1, 2] was one of the first methods to introduce the possibility of measuring the semantic similarity when comparing a text written by a student to the corresponding learning base. Later on, Latent Dirichlet Allocation (LDA) [3] was introduced as a topic modeling technique that overcomes some problems

E. André et al. (Eds.): AIED 2017, LNAI 10331, pp. 52–63, 2017.
DOI: 10.1007/978-3-319-61425-0_5

of LSA. Even if LSA and LDA are powerful techniques, due to their inherited bag of words approach, they cannot be used alone for evaluating the complexity and quality of a written text.

Our aim is to build a comprehensive Automated Essay Scoring model for Dutch language. However, text complexity is a hard to define concept and, therefore, it cannot be measured with only a few metrics. Moreover, the complexity of a text is directly related to its ease of reading and to comprehension, which means it also involves human reader particularities, for example, age, level of knowledge, socio-cultural features, and even skill and motivation. *Coherence*, the main feature of a good discourse, of a good quality text, a premise of reducing complexity, is also related to human's perception and it is very hard to measure [4]. *Cohesion* is a simpler to handle and operationalize concept that is tightly connected to semantic similarity.

Many metrics and qualitative criteria for analyzing complexity have been proposed, as it will be discussed in the next section, and various computer systems for computing such metrics have become available [5]. In the research presented in this paper, we used the *ReaderBench* NLP framework [6, 7], which integrates a wide range of metrics and techniques, covering both the cognitive and socio-cultural paradigms. *ReaderBench* makes extensive usage of Cohesion Network Analysis (CNA) [8, 9] in order to represent discourse in terms of semantic links; this enables the computation of various local and global cohesion measures described later on. In addition, *ReaderBench* is grounded in Bakhtin's dialogism [10], which provides a unified framing for both individual and collaborative learning [9, 11].

An important parameter that should be considered for AES is the specific language. First, LSA, LDA and any statistical approaches for analyzing essays require text corpora written in the language of the essays. Second, there may be significant differences among languages with respect to the average length of sentences and even words, size of vocabulary, discourse structuring, etc. Dutch language, in contrast to English, contains a high number of compound words (which inherently decreases the number of tokens per phase); moreover, besides compound words, general words tend to be longer [12]. In this idea, this paper presents the stages required for porting the *ReaderBench* framework, which was developed mainly for English, to Dutch language.

The paper continues with a state of the art section, followed by an in-depth presentation of the undergone steps required to build our comprehensive Dutch assessment model. Our evaluation is based on a corpus of student reports in the domain of environmental sciences. While engaging in a serious game, students adopt the role of principal researcher for investigating a multifaceted environmental problem and, on various occasions throughout the game. they are required to report about their findings. After discussing the results, the fifth section presents the conclusions, as well as further enhancements to be integrated within our approach.

2 State of the Art

The idea of quantifying textual complexity or difficulty has been studied intensively over the years, having in mind two major goals: presenting readers with materials aligned with their level of comprehension, and evaluating learners' abilities and

knowledge levels from their writing traces. In our current research, we are focusing on the latter goal, evaluating students' writing capabilities in order to discover significant correlations to their knowledge level.

From a global perspective, textual complexity is relative to the student's knowledge of the domain, language familiarity, interest and personal motivation [6]. In addition, the reader's education, cognitive capabilities and prior experiences influence readability and comprehension [6]. In accordance to the Common Core State Standards Initiative [13], textual complexity can be evaluated from three different perspectives: *quantitative* (e.g., word frequency, word/phrase length), *qualitative* (e.g., clarity, structure, language familiarity) and from the *reader and task orientation* (e.g., motivation, prior knowledge or interest). In practice, these dimensions of textual complexity can be used to determine if a student is prepared for college or for a career. The scope of the standard is to reduce and eliminate knowledge gaps by offering students a coherent flow of materials that have a slightly higher textual complexity in order to challenge the reader.

A significant effort has been put into developing automated tools of textual complexity assessment as part of the linguistic research domain. *E-Rater* [14] is one of the first automated systems to evaluate text difficulty based on three general classes of essay features: structure (e.g., sentence syntax, proportion of spelling, grammar, usage or mechanics errors), organization based on various discourse features, and content based on prompt-specific vocabulary. Several other tools for automated essay grading or for assessing the textual complexity of a given text have been developed and employed in various educational programs [5, 15]: *Lexile* (MetaMetrics), *ATOS* (Renaissance Learning), *Degrees of Reading Power: DRP Analyzer* (Questar Assessment, Inc.), *REAP* (Carnegie Mellon University), *SourceRater* (Educational Testing Service), *Coh-Metrix* (University of Memphis), Markit (Curtin University of Technology) [16], IntelliMetric [17] or *Writing Pal* (Arizona State University) [18, 19].

In terms of Dutch language, there are only a few systems that perform automated essay scoring by integrating multiple textual complexity indices. T-Scan (http://languagelink.let.uu.nl/tscan) is one of the most elaborated solutions as it considers multiple features, including [20]: lexical and sentence complexity, referential cohesion and lexical diversity, relational coherence, concreteness, personal style, verbs and time, verbs and time, as well as probability features, all derived from Coh-Metrix [21–23]. Besides T-Scan, various Dutch surface tools have been reported that provide lexical indices for text difficulty, as well as recommendations to reorganize the text: e.g., Texamen, Klinkende Taal and Accessibility Leesniveau Tool [24].

3 Building the Dutch Complexity Model

3.1 The NLP Processing Pipeline for Dutch Language

Before establishing a comprehensive list of textual complexity indices that can be used to predict a learner's understanding level, we first need to build a Natural Language Processing (NLP) pipeline for Dutch language. This processing pipeline integrates key techniques that are later on used also within the scoring algorithm. Multiple challenges

were encountered besides mere translation issues while adapting our *ReaderBench* framework from English to Dutch language; thus, we see fit to provide prescriptive information regarding our NLP specific processes.

First, a new thorough dictionary was required to perform a comprehensive cleaning of the input text, by filtering and selecting only dictionary words. Elimination of noise within the unsupervised training process of semantic models, as well as facile identification of typos are important elements while building our textual complexity model. Moreover, as the essays used were academic reports we were also constrained to include low-frequency, scientific words, in order to be capable to grasp the specificity of our texts. E-Lex (formerly named TST-lexicon) [25] is a lexical database of Dutch language consisting of both one-word and multi-word lexicons, and it represented the best starting point after manually reviewing multiple dictionaries. Besides providing a comprehensive list of words, E-Lex was also used to build a static lemmatizer that reduces each inflected word form to its corresponding lemma, therefore normalizing the input.

Second, similar to the requirement of a new dictionary, a new stop words list (i.e. words having limited or no content information) was required in order to disregard certain words for scoring purposes. Again, upon manual review, we opted for http://snowball.tartarus.org/algorithms/dutch/stop.txt which was expanded with numbers, interjections, as well frequent words with low semantic meaning. These words induced noise within the emerging topics from Latent Dirichlet Allocation (LDA) [3] by having a high occurrence rate, as well as a high probability, in multiple topics.

Third, new semantic models, namely vector space models based on Latent Semantic Analysis [1] and Latent Dirichlet Allocation topic distributions [3] needed to be trained. The Corpus of Contemporary Dutch (Hedendaags Nederlands; 1.35 billion words; http://corpushedendaagsnederlands.inl.nl) represented the best alternative in terms of dimension, breadth of topics, as well as novelty of comprised documents. After preprocessing, the corpus was reduced to around 500 million content words from approximately 11.5 million paragraphs, each surpassing the minimum imposed threshold of at least 20 content words. The LSA space was built using the stochastic SVD decomposition from Apache Mahout [26] which was applied on the term-document matrix weighted with log-entropy, across 300 dimensions. LDA made use of parallel Gibbs sampling implemented in Mallet [27] and the model was created with 100 topics, as suggested by Blei [28]. A manual inspection of top 100 words from each LDA topic suggested that the space was adequately constructed due to the fact that the most representative words from each topic were semantically related one to another.

Fourth, complementary to our LSA and LDA models, the Open Dutch WordNet, the most complete Dutch lexical semantic database up-to-date with more than 115,000 synsets, was also integrated, enabling the following: (a) the identification of lexical chains and word sense disambiguation [29], as well as (b) the computation of various semantic distances in ontologies, namely Wu-Palmer, Leacock-Chodorow and path length distances [30].

3.2 Textual Complexity Indices

Starting from the wide range of textual complexity indices available within the *ReaderBench* framework [6, 7] for English language, and based on the previously described NLP processing pipeline, we present the multitude of textual complexity indices that we have made available into Dutch language.

In contrast to the systems mentioned within the state of the art section and besides covering multiple layers of the analysis ranging from surface indices, syntax to semantics, *ReaderBench* focuses on text cohesion and discourse connectivity. The framework provides a more in-depth perspective of discourse structure based on Cohesion Network Analysis [8, 9], a multi-layered cohesion graph [31] that considers semantic links between different text constituents. We further describe the indices integrated in our framework and used for this study, categorized by their textual analysis scope.

Surface, lexicon and syntax analyses. The first approaches to text complexity were developed by Page [32] in his search to develop an automatic grading system for students' essays. Page discovered a strong correlation between human intrinsic variables (trins) and proxes (i.e., computer approximations or textual complexity indices), thus proving that statistical analyses can provide reliable textual automated estimations. Our model integrates the most representative and predictive proxes from Page's initial study, corroborated with other surface measures frequently used in other automated essay grading systems (e.g., average word/phrase/paragraph length, average unique/content words per paragraph, average commas per sentence/paragraph). Entropy at word level, derived from Shannon's Information Theory [33], is a relevant metric for quantifying textual complexity based on the hypothesis that a more complex text contains more information, more diverse concepts and requires more working memory. In contrast, character entropy is a language specific characteristic [34] and does not exhibit a significant variance in texts written in English. Moreover, of particular interest at this level due to the inherit implications in co-reference resolution, are the different categories of pronouns (i.e., first, second and third person, interrogative, and indefinite pronouns), implemented as predefined words lists and considered within our model. Coverage statistics with regards to specific pronouns usage were computed at sentence, paragraph, and document levels.

Semantic analysis and discourse structure. In order to comprehend a text, the reader must create a coherent and well connected representation of the information, commonly referred to as the situation model [35]. According to McNamara et al. [15], textual complexity is linked with cohesion in terms of comprehension, as the lack of cohesion can artificially increase the perceived difficulty of a text. Thus, our model uses a local and global evaluation of cohesion within the CNA graph, computed as the average value of the semantic similarities of all linksat intra- and inter-paragraph levels [31, 36]. Cohesion is estimated as the average value of [6]: (a) Wu-Palmer semantic distances applied on the WordNet lexicalized ontology, (b) cosine similarity in Latent Semantic Analysis (LSA) vector space models, and (c) the inverse of the Jensen Shannon dissimilarity (JSD) between Latent Dirichlet Allocation (LDA) topic distributions [37].

Besides semantic models, lexical chains provide a strong basis for assessing text cohesion and several indices have been also introduced: (a) the average and the maximum span of lexical chains (the distance in words between the first and the last occurrence of words pertaining to the same chain), (b) the average number of lexical chains per paragraph, as well as (c) the percentage of words that are included in lexical chains (i.e., words that are not isolated within the discourse, but inter-linked with other concepts from the same chain).

In addition, starting from the Referentiebestand Nederlands (RBN) [38], several discourse connectors identifiable via cue phrases have been added to our complexity model in order to provide a fine-grained view over the discourse with regards to the following relevant relationships: cause, circumstance, comparison, concession, condition, conjunctive, contrast, degree, disjunctive, effect, exception, nonrestrictive, other, purpose, restriction, time, and interrogative.

Word complexity represents a mixture of different layers of discourse analysis covering a wide set of estimators for each word's difficulty: (a) syllable count, (b) distance in characters between the inflected form, lemma and word stem (adding multiple prefixes or suffixes increases the difficulty of using a certain word), (c) specificity reflected in the inverse document frequency from LSA/LDA training corpus, (d) the average and the maximum path distance in the hypernym tree based on all word senses and (e) the word polysemy count from WordNet [39]. In order to reflect individual scores at sentence and paragraph level, all these indices were averaged, taking into consideration only lemmatized content words generated after applying the NLP processing pipeline. Moreover, normalized occurrences at both paragraph and sentence levels of all major word categories from the Dutch LIWC dictionary [40] have been considered, providing additional insights in terms of underlying concept categories.

3.3 Automated Text Segmentation

The previously introduced textual complexity indices become less relevant when facing longer documents comprising of thousands or tens of thousands of words. Besides the computational power required for building a complete CNA graph that captures all potential cohesive links, different sections might exhibit different traits which can be easily disregarded at document level. A commonly encountered approach is to automatically split longer texts using an imposed fixed window of words. The most frequently used threshold value is of 1,000 words [5]. However, this method fails to consider the natural discourse structure of the text, its hierarchical decomposition, as most documents contain sections, subsections and so forth, constituent elements that emerge as a more viable manner of splitting the text. Therefore, the headings from the initial document produce a hierarchical structure in which each section contains its own text and list of subsections that can be possibly empty.

Thus, we developed a new segmentation method applicable for Microsoft Word documents, assuming that sections are correctly annotated with the appropriate heading styles reflecting its hierarchical structure (e.g., Heading 1 is automatically considered as a section, Heading 2 a subsection, Heading 3 a subsubsection, etc.). From a technical perspective, due to the constraint that the entire framework is written entirely in Java,

we have opted to rely on the Apache POI library (https://poi.apache.org) for parsing the *.docx* documents. The newly generated meta-document contains multiple layers of well-defined and self-contained document segments on which we can apply the previously introduced textual complexity indices. The results for each textual complexity index and for each extracted section are averaged in order to obtain the scores for the entire meta-document.

4 Results

4.1 Corpus

The corpus used for performing a preliminary validation of our model consisted of 173 technical reports in Dutch written by master degree students from the Open University of the Netherlands and Utrecht University. The students play an online game in the domain of environmental policy, which confronts them multidimensional environmental problems. During the game, they are required to upload technical reports about their findings, in subsequent stages (i.e., analysis, 2 design tasks, 2 evaluation tasks and a final evaluation) [41]. As these reports need to be evaluated manually by teachers in very short time spans, the need for Automated Essay Scoring arose. All essays are scored by human tutors on the bases of an assessment framework and scores express a linear variable ranging from 1 (utterly weak) to 10 (excellent). The reports used for this experiment address only the first stage (i.e., analysis) and contained an average of 1832 words (SD = 790), ranging from a minimum of 243 words to a maximum of 6186 words. All reports were manually corrected in terms of formatting in order to ensure an appropriate usage of heading styles, a process that afterwards facilitates their automated assessment.

Because of the limited number of students whose scores span multiple levels, we applied a binary split of student scores into two distinct classes: high performance students with scores ≥ 7, while the rest were catalogued as low performance students. Moreover, for the scope of these preliminary experiments, we opted to rely only on the LDA topic model besides WordNet, instead of both LSA and LDA. This was due to the fact that only the LDA space was inspected by native speakers with regards to comprising relevantword associations within corresponding topics.

4.2 Statistical Analyses

The Dutch indices from *ReaderBench* that lacked normal distributions were discarded (e.g., average number of sentences, words and content words, average number of commas at paragraphs and sentence levels, word polysemy counts, different connectors and word lists at paragraph and sentence level). Correlations between the selected indices and the dependent variable (the students' score for their technical report) were then calculated for the remaining indices to determine whether there was a statistically significant relation ($p < .05$). Indices that were highly collinear ($r \geq .9$) were flagged, and the index with the strongest correlation with the assigned score corresponding to

each report was retained, while the other indices were removed. The remaining indices were included as predictor variables in a stepwise regression to explain the variance in the students' scores, as well as predictors in a Discriminant Function Analysis [42] used to classify students based on their performance.

4.3 Relationship Between *ReaderBench* and Students' Final Scores

To address our research question of automatically scoring students' reports, we conducted correlations between the *ReaderBench* indices that were normally distributed and were not multicollinear and their final scores. As shown in Table 1, medium to weak effects were found for *ReaderBench* indices related to the number of words, paragraphs, unique words per sentence, lexical chains, lower local cohesion induced by a more varied vocabulary (higher word entropy), different types of discourse connectors at both sentence and paragraph levels (concession, condition, circumstance), as well as pronouns (both third person and indefinite).

Table 1. Correlations between *ReaderBench* indices and report score.

Index	*r*	*p*
Logarithmic number of words	.461	<.001
Average number of lexical chains per paragraph	.338	<.001
Average sentence-paragraph cohesion (Wu-Palmer semantic distance in WordNet)	-.284	<.001
Average number of concession connectors per paragraph	.269	<.001
Average number of condition connectors per paragraph	.260	.001
Word entropy	.258	.001
Average number of circumstance connectors per paragraph	.254	.001
Percentage of words that are included in lexical chains	.250	.001
Average number of indefinite pronouns per sentence	.237	.002
Average sentence length (number of characters)	.193	.011
Average number of third person pronouns per sentence	.187	.014
Average number of circumstance connectors per sentence	.187	.014
Average number of unique content words per sentence	.184	.015
Number of paragraphs	.160	.035
Average number of condition connectors per sentence	.154	.044

The correlations indicate that students who received higher scores had longer reports in terms of words and paragraphs, greater word entropy, used more discourse connectors and pronouns, and produced more unique words. Moreover, students who received higher scores had lower inner cohesion per paragraph, indicating more elaborated paragraphs that reflect a mixture of diverse ideas.

4.4 Regression Analysis and Discriminant Function Analysis

To analyze which *ReaderBench* features best predicted the students' score, we conducted a stepwise regression analysis using the 15 significant indices as the independent variables. This yielded a significant model, $F(3, 169) = 24.676$, $p < .001$, $r = .552$, $R^2 = .305$. Three variables were significant and positive predictors of report scores: logarithmic number of words, average number of pronouns per sentence (indefinite), percentage of words that are included in lexical chains. These variables explained 30.5% of the variance in the students' report scores.

The stepwise Discriminant Function Analysis (DFA) retained three different variables as significant predictors (i.e., 1. logarithmic number of words, 2. average number of indefinite pronouns per sentence, and 3. average sentence-paragraph cohesion using Wu-Palmer semantic distance), and removed the remaining variables as non-significant predictors.

Table 2. Confusion matrix for DFA classifying students based on performance

		Predicted performance membership		Total
		Low	High	
Whole set	Low	54	21	75
	High	20	78	98
Cross-validated	Low	53	22	71
	High	21	77	98

The results prove that the DFA using these three indices correctly allocated 132 of the 173 students from our dataset, $\chi^2(df = 3, n = 173) = 40.948$, $p < .001$, for an accuracy of 76.3% (the chance level for this analysis is 50%). For the leave-one-out cross-validation (LOOCV), the discriminant analysis allocated 130 of the 173 students for an accuracy of 75.1% (see the confusion matrix reported in Table 2 for results). The measure of agreement between the actual student performance and that assigned by our model produced a weighted Cohen's Kappa of .517, demonstrating moderate agreement.

5 Conclusions

The *ReaderBench* NLP framework was extended to support automatic scoring of students' technical reports written in Dutch language. Existing textual complexity indices and methods had to be adapted from English language, and specifically tweaked for Dutch language, thus introducing one of the most comprehensive models available for Dutch to our knowing. Moreover, we have also introduced an automatic segmentation method that creates a hierarchical structure based on document sections and headings.

Initial results indicate that our model, which goes beyond the replication of the English version of *ReaderBench* due to the performed customizations, has a high accuracy and is suitable for automatically scoring Dutch technical reports. In addition, the performance of our model is comparable to systems available in English language. Our framework integrates the widest range of textual complexity indices available for Dutch language, emphasizing the semantic dimension of the analysis instead of frequently used surface measures. Nevertheless, we must point out that the variance explained by the regression model, as well as the weighted Cohen's Kappa, are rather low in contrast to the accuracy of the DFA model which only assumes a binary classification. Only the index with the highest correlation (i.e., logarithmic number of words) was retained in both the linear regression and in the DFA model. The remaining indices are specific for each model that is fundamentally different – the regression model predicts a linear score, while the DFA performs a classification into two performance categories.

As limitations, we must also point out the discrepancies in the evaluation of the technical reports as the automatic evaluation is mostly focused on students' writing style, while the tutors evaluate the technical quality of the report. Moreover, the population for our study consists of master degree students who have, in general, relatively high writing skills; in return, this may reduce the variance in complexity among the essays. Therefore, new metrics should be introduced in order to address the technical soundness of a document in relation to a given theme or an imposed set of topics of interest. Moreover, the Dutch language imposes additional challenges, like the high number of compound words. While relating to the process of building semantic models, these words could be more relevant if taken separately. Thus, automated splitting rules should be enforced upon compound words in order to provide a clearer contextualization of the input text.

Acknowledgments. This work was partially funded by the 644187 EC H2020 *Realising an Applied Gaming Eco-system* (RAGE) project, by the FP7 208-212578 LTfLL project, as well as by University Politehnica of Bucharest through the "Excellence Research Grants" Program UPB–GEX 12/26.09.2016.

References

1. Landauer, T.K., Dumais, S.T.: A solution to Plato's problem: the Latent Semantic Analysis theory of acquisition, induction and representation of knowledge. Psychol. Rev. **104**(2), 211–240 (1997)
2. Miller, T.: Essay assessment with Latent Semantic Analysis. J. Educ. Comput. Res. **29**(4), 495–512 (2003)
3. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. J. Mach. Learn. Res. **3**(4–5), 993–1022 (2003)
4. Crossley, S.A., McNamara, D.S.: Text coherence and judgments of essay quality: models of quality and coherence. In: 33rd Annual Conference of the Cognitive Science Society, pp. 1236–1231. Cognitive Science Society, Boston (2011)

5. Nelson, J., Perfetti, C., Liben, D., Liben, M.: Measures of Text Difficulty: Testing their Predictive Value for Grade Levels and Student Performance. Council of Chief State School Officers, Washington, DC (2012)
6. Dascalu, M.: Analyzing Discourse and text complexity for learning and collaborating, Studies in Computational Intelligence, vol. 534. Springer, Cham (2014)
7. Dascalu, M., Dessus, P., Bianco, M., Trausan-Matu, S., Nardy, A.: Mining texts, learner productions and strategies with *ReaderBench*. In: Peña-Ayala, A. (ed.) Educational Data Mining. SCI, vol. 524, pp. 345–377. Springer, Cham (2014). doi:10.1007/978-3-319-02738-8_13
8. Dascalu, M., McNamara, D.S., Trausan-Matu, S., Stavarache, L.L., Allen, L.K.: Cohesion network analysis of CSCL participation. Behavior Research Methods, PP. 1–16 (2017)
9. Dascalu, M., Trausan-Matu, S., McNamara, D.S., Dessus, P.: ReaderBench – automated evaluation of collaboration based on cohesion and dialogism. Int. J. Comput. Support. Collaborative Learn. **10**(4), 395–423 (2015)
10. Bakhtin, M.M.: The dialogic imagination: four essays. The University of Texas Press, Austin (1981)
11. Dascalu, M., Allen, K.A., McNamara, D.S., Trausan-Matu, S., Crossley, S.A.: Modeling comprehension processes via automated analyses of dialogism. In: 39th Annual Meeting of the Cognitive Science Society (CogSci 2017). Cognitive Science Society, London (2017, in Press)
12. Duyck, W., Desmet, T., Verbeke, L.P., Brysbaert, M.: WordGen: A tool for word selection and nonword generation in Dutch, English, German, and French. Behav. Res. Methods **36**(3), 488–499 (2004)
13. National Governors Association Center for Best Practices & Council of Chief State School Officers: Common Core State Standards. Authors, Washington D.C. (2010)
14. Powers, D.E., Burstein, J., Chodorow, M., Fowles, M.E., Kukich, K.: Stumping e-rater®: Challenging the Validity of Automated Essay Scoring. Educational Testing Service, Princeton (2001)
15. McNamara, D.S., Graesser, A.C., Louwerse, M.M.: Sources of text difficulty: Across the ages and genres. In: Sabatini, J.P., Albro, E., O'Reilly, T. (eds.) Measuring up: Advances in How we Assess Reading Ability, pp. 89–116. R&L Education, Lanham (2012)
16. Williams, R., Dreher, H.: Automatically grading essays with Markit©. J. Issues Informing Sci. Inform. Technol. **1**, 693–700 (2004)
17. Elliot, S.: IntelliMetric: from here to validity. In: Shermis, M.D., Burstein, J.C. (eds.) Automated Essay Scoring: A Cross Disciplinary Approach, pp. 71–86. Lawrence Erlbaum Associates, Mahwah (2003)
18. Crossley, S.A., Allen, L.K., McNamara, D.S.: The Writing Pal: a writing strategy tutor. In: Crossley, S.A., McNamara, D.S. (eds.) Handbook on Educational Technologies for Literacy. Taylor & Francis, Routledge, New York (in press)
19. McNamara, D.S., Crossley, S.A., Roscoe, R., Allen, L.K., Dai, J.: A hierarchical classification approach to automated essay scoring. Assessing Writ. **23**, 35–59 (2015)
20. Pander Maat, H.L.W., Kraf, R.L., van den Bosch, A., van Gompel, M., Kleijn, S., Sanders, T.J.M., van der Sloot, K.: T-Scan: a new tool for analyzing Dutch text. Comput. Linguist. Neth. J. **4**, 53–74 (2014)
21. Graesser, A.C., McNamara, D.S., Louwerse, M.M., Cai, Z.: Coh-Metrix: Analysis of text on cohesion and language. Behav. Res. Methods Instrum. Comput. **36**(2), 193–202 (2004)
22. Graesser, A.C., McNamara, D.S., Kulikowich, J.M.: Coh-Metrix: Providing multilevel analyses of text characteristics. Educ. Res. **40**(5), 223–234 (2011)
23. McNamara, D.S., Graesser, A.C., McCarthy, P., Cai, Z.: Automated Evaluation of Text and Discourse with Coh-Metrix. Cambridge University Press, Cambridge (2014)

24. Kraf, R., Lentz, L., Pander Maat, H.: Drie Nederlandse instrumenten voor het automatisch voorspellen van begrijpelijkheid. Een klein consumentenonderzoek. Tijdschift voor Taalbeheersing **33**(3), 249–265 (2011)
25. CGN Consortium: e-Lex, lexicale databank (lexical database). Instituut voor Nederlandse Taal, Leiden, the Netherlands (2017)
26. Owen, S., Anil, R., Dunning, T., Friedman, E.: Mahout in Action. Manning Publications Co., Greenwich (2011)
27. McCallum, A.K.: MALLET: A Machine Learning for Language Toolkit (2002). http://mallet.cs.umass.edu/
28. Blei, D.M.: Probabilistic topic models. Commun. ACM **55**(4), 77–84 (2012)
29. Galley, M., McKeown, K.: Improving word sense disambiguation in lexical chaining. In: 18th International Joint Conference on Artificial Intelligence (IJCAI 2003), pp. 1486–1488. Morgan Kaufmann Publishers, Inc., Acapulco (2003)
30. Budanitsky, A., Hirst, G.: Evaluating WordNet-based measures of lexical semantic relatedness. Comput. Linguist. **32**(1), 13–47 (2006)
31. Trausan-Matu, S., Dascalu, M., Dessus, P.: Textual complexity and discourse structure in computer-supported collaborative learning. In: Cerri, Stefano A., Clancey, William J., Papadourakis, G., Panourgia, K. (eds.) ITS 2012. LNCS, vol. 7315, pp. 352–357. Springer, Heidelberg (2012). doi:10.1007/978-3-642-30950-2_46
32. Wresch, W.: The imminence of grading essays by computer—25 years later. Comput. Compos. **10**(2), 45–58 (1993)
33. Shannon, C.E.: Prediction and entropy of printed English. Bell Syst. Tech. J. **30**, 50–64 (1951)
34. Gervasi, V., Ambriola, V.: Quantitative assessment of textual complexity. In: Barbaresi, M.L. (ed.) Complexity in Language and Text, pp. 197–228. Plus, Pisa, Italy (2002)
35. van Dijk, T.A., Kintsch, W.: Strategies of Discourse Comprehension. Academic Press, New York (1983)
36. Dascalu, M., Dessus, P., Trausan-Matu, Ş., Bianco, M., Nardy, A.: *ReaderBench*, an environment for analyzing text complexity and reading strategies. In: Lane, H.Chad, Yacef, K., Mostow, J., Pavlik, P. (eds.) AIED 2013. LNCS, vol. 7926, pp. 379–388. Springer, Heidelberg (2013). doi:10.1007/978-3-642-39112-5_39
37. Manning, C.D., Schütze, H.: Foundations of statistical Natural Language Processing. MIT Press, Cambridge (1999)
38. van der Vliet, H.: The Referentiebestand Nederlands as a multi-purpose lexical database. Int. J. Lexicogr. **20**(3), 239–257 (2007)
39. Miller, G.A.: WordNet: a lexical database for English. Commun. ACM **38**(11), 39–41 (1995)
40. Zijlstra, H., van Meerveld, T., van Middendorp, H., Pennebaker, J.W., Geenen, R.: De Nederlandse versie van de Linguistic Inquiry and Word Count (LIWC), een gecomputeriseerd tekstanalyseprogramma [Dutch version of the Linguistic Inquiry and Word Count (LIWC), a computerized text analysis program]. Gedrag & Gezondheid **32**, 273–283 (2004)
41. Westera, W., Nadolski, N., Hummel, H.: Serious gaming analytics: what students' log files tell us about gaming and learning. Int. J. Serious Games **1**(2), 35–50 (2014)
42. Klecka, W.R.: Discriminant Analysis. Quantitative Applications in the Social Sciences Series, vol. 19. Sage Publications, Thousand Oaks (1980)